Information Retrieval Journal (2019) 22:476–498
https://doi.org/10.1007/s10791-018-09350-9

On the impact of group size on collaborative search effectiveness

Felipe Moraes[1] · **Kilian Grashoff**[1] · **Claudia Hauff**[1]

Received: 16 August 2018 / Accepted: 18 December 2018 / Published online: 9 January 2019

Abstract
While today's web search engines are designed for single-user search, over the years research efforts have shown that complex information needs—which are explorative, open-ended and multi-faceted—can be answered more efficiently and effectively when searching *in collaboration*. Collaborative search (and sensemaking) research has investigated techniques, algorithms and interface affordances to gain insights and improve the collaborative search process. It is not hard to imagine that the *size* of the group collaborating on a search task significantly influences the group's behaviour and search effectiveness. However, a common denominator across almost all existing studies is a fixed group size—usually either pairs, triads or in a few cases four users collaborating. Investigations into larger group sizes and the impact of group size dynamics on users' behaviour and search metrics have so far rarely been considered—and when, then only in a simulation setup. In this work, we investigate in a large-scale user experiment to what extent those simulation results carry over to the real world. To this end, we extended our collaborative search framework `SearchX` with algorithmic mediation features and ran a large-scale experiment with more than 300 crowd-workers. We consider the collaboration group size as a dependent variable, and investigate collaborations between groups of up to six people. We find that most prior simulation-based results on the impact of collaboration group size on behaviour and search effectiveness *cannot* be reproduced in our user experiment.

Keywords Collaborative search · Search effectiveness · Interactive search

✉ Claudia Hauff
c.hauff@tudelft.nl

Felipe Moraes
f.moraes@tudelft.nl

Kilian Grashoff
k.c.grashoff@student.tudelft.nl

[1] Delft University of Technology, Delft, The Netherlands

1 Introduction

Collaborative search and more generally online collaborative information seeking have been shown to be effective tools to tackle complex information needs, i.e., information needs that are explorative, open-ended and multi-faceted (Shah 2010). In this setting, all collaborators have the same information need and are aware of each other's activities. Those information needs do not only occur in information-intensive work domains such as the patent domain (Hansen and Järvelin 2005) but also in personal search scenarios such as travel planning (Morris 2008; Kelly and Payne 2014), personal health (Morris 2013), and online shopping (Gao et al. 2016).

Collaborative search frameworks and approaches are commonly categorised along four dimensions (Golovchinsky et al. 2009):

- intent (explicit vs. implicit collaboration),
- depth of mediation (interface-based vs. algorithmic or hybrid mediation),
- concurrency (asynchronous vs. synchronous search), and
- location (co-located vs. remote).

In this work, we focus on *remote* and *explicit collaborations*, that is, collaborations between users who share a common information need and each work on their own device—this is in contrast to implicit collaborations where no shared information need exists and users' traces are leveraged to personalise search, e.g. (Teevan et al. 2009; Smyth et al. 2004), and co-located collaborations with several users collaborating on a single device such as a TableTop (Smeaton et al. 2007).

One dimension that has largely been considered as a constant in explicit collaborative search studies is the *size* of the collaborating group. Almost all existing collaborative search studies—no matter if evaluated in a lab or a simulation setting—consider groups of either **two** (Brennan et al. 2008; Kelly and Payne 2014; Morris and Horvitz 2007; Pickens et al. 2008; Joho et al. 2008; Shah et al. 2010; Soulier et al. 2014a, b; Tamine and Soulier 2015; Htun et al. 2015; Böhm et al. 2016; Htun et al. 2017; González-Ibáñez et al. 2013; Shah and González-Ibáñez 2011), **three** (Amershi and Morris 2008; Morris et al. 2008) or **four** (Capra et al. 2012; Paul and Morris 2009) users. This strong focus on pairs of collaborators can be explained by the fact that many studies investigate novel collaborative search features to, for instance, increase awareness of group members' interactions; to facilitate the sharing of knowledge among users; and to employ division of labour. As the number of experimental variants in such studies can increase quickly, a common way to limit the number of variants is to keep the group size constant. In addition, fixing the group size to either two or three users is often motivated by two large-scale surveys that were conducted on the use of collaborative search in 2006 and 2012 respectively (Morris 2008, 2013). During that time period, the frequency of collaborative search episodes increased ten-fold with more than 10% of surveyed users in 2012 reporting daily collaborative search episodes. When asked about their most recent collaborative search episode (in the 2012 study), slightly more than half of the participants reported this having been a collaboration between two or three users; at the same time, more than 21% of respondents reported group sizes of five and more. Thus, collaborations between more than three users are not a rare occurrence, though they are hardly investigated in research.

An exception to the observations above is (Joho et al. 2009) who—in a simulation study—investigated the effect of *changing* group sizes in collaborative search. Groups of

up to five *simulated* users were explored across a range of algorithmic mediation strategies (some of which we also explore in our work); the authors found larger group sizes to lead to higher search effectiveness in a recall-oriented task, albeit with diminishing returns. Simulations though are by definition simplifications of the real world. They ignore the increase in cognitive load real users are likely to experience as group sizes increase and coordination efforts become more difficult to manage. It is thus an open question (1) to what extent the simulation findings carry over to a real user study, and, (2) to what extent the currently existing collaborative search mechanisms (algorithms as well as interface elements) *scale* to group sizes beyond the commonly investigated sizes of two or three.

In order to investigate these issues, we extended our collaborative search framework `SearchX` (Putra et al. 2018) with algorithmic mediation (among others, shared relevance feedback) that was found to be effective in prior works. We designed a crowd-sourcing study with 305 participants based on prior best principles with group size as main dependent variable, investigating groups of 2, 4 and 6 collaborating searchers. The analyses we present here are guided by the following overarching research question: **What is the impact of group size on collaborative search effectiveness?**

The main contributions and findings of our work are:

- We extended `SearchX`, our synchronous collaborative search framework with algorithmic mediation components as well as features enabling efficient use of `SearchX` for crowdsourcing studies, that we successfully deployed in a crowd-sourcing setup with hundreds of crowd-workers and different levels of user and task synchronisation.
- We find most prior simulation-based results on the impact of group size on behaviour and search effectiveness to not hold in our user study with several hundred crowd-workers.
- Importantly, in our study—conducted across three difficult recall-oriented search topics—we do *not* observe diminishing returns (measured in recall) when scaling up group sizes from two to six collaborators. Our results indicate that a further scaling up of group sizes is feasible with existing collaborative search features and can potentially lead to new research avenues in the collaborative search space.

2 Related work

A *collaboration* is a *"true synergy among diverse participants in creating solutions or strategies through the synergistic interactions of a group of people"* (Kapetanios 2008). In the search setting, previous work has shown that users often engage in collaborative search activities (Morris 2008, 2013; Twidale et al. 1997) when dealing with complex information needs, using search tools designed for single-user search (web search engines being the most prominent example). When introduced to dedicated collaborative search systems users consider them usable and appropriate for collaborative search tasks (Kelly and Payne 2014). While Shah and González-Ibáñez (2011) showed empirically that a *synergy* effect holds in collaborative search (pairs of collaborating searchers being significantly more effective than pairs of independent searchers), no change in search effectiveness between pairs of independent and collaborating searchers was observed in Joho et al. (2008).

For the collaborative search setting we consider in this work (*remote* and *explicit*), four design principles have been formulated by Morris (2007):

Table 1 Overview of key statistics of empirical evaluations of collaborative search: group size (GS), number of groups (#G), number of search tasks per group (#T) and study type: [sim.] refers to a simulation study with batch evaluation, [lab-fixed] to a lab user study with one or more fixed work/personal search tasks, [lab-nat.] to a lab user study where users self-selected their search task(s)

	GS	#G	#T	Type	Collection
Morris and Horvitz (2007)	2	7	1	Lab-nat.	Web
Amershi and Morris (2008)	3	12	3	Lab-nat.	Web
Morris et al. (2008)	3	10	1	Lab-nat.	Web
Pickens et al. (2008)	2	4	24	Lab-fixed	TRECVid07
Joho et al. (2008)	2	12	3	Lab-fixed	Aquaint
Paul and Morris (2009)	4	12	1	Lab-fixed	Web
Joho et al. (2009)	1–5	500	13	Sim.	Aquaint
Shah et al. (2010)	2	5	10	Sim.	-
Capra et al. (2012)	4	11	1	Lab-fixed	Aquaint
González-Ibáñez et al. (2013)	2	30	1	Lab-fixed	Web
Kelly and Payne (2014)	2	8	1–3	Lab-nat.	Web
Soulier et al. (2014b)	2	–	20	Sim.	TREC Vol. 4
Soulier et al. (2014a)	2	70	1	Lab-fixed	Web
Tamine and Soulier (2015)	2	75	1	Lab-fixed	Web
Htun et al. (2015)	2	55	13	Sim.	Aquaint
Böhm et al. (2016)	2	–	314	Sim.	OHSUMED, CLEF-IP
Htun et al. (2017)	2	10	3	Lab-fixed	Aquaint
Our work	2-6	67	3	Lab-fixed	Aquaint

Collection refers to the data collection used

– indicates unknown

1. *raising awareness* among the collaborators about their activities (e.g. by providing a shared query history);
2. enabling the *division of labour* (e.g. by automatically providing different results to collaborators or enabling collaborators to chat and divide the search task manually);
3. *persistence* (e.g. by storing the query history persistently);
4. enabling *sensemaking* (e.g. by providing multiple views of the common activities).

The first two design principles are most often explored in the information retrieval community and approaches to raising awareness and dividing the labour can be categorized as belonging to one of three levels (Joho et al. 2009): the *interface level* (e.g. a chat widget enables users to manually divide the work), the *techniques level* (i.e. established IR technologies such as document clustering are employed to facilitate the collaboration) or the *ranking model level* (i.e. the ranking model is adapted specifically for the collaborative use case).

We now present prior works in each of these categories in turn and then finish the section with an overview of existing tools and a detailed look at prior works on group size dynamics in collaborative search. We focus in particular on the ideas presented in prior works, as the reported findings are often based on small user studies—some of those more than 10 years old—that explore the use of a single collaborative search system in a single setting. Table 1 presents an overview of key user study statistics (group sizes investigated, number of topics,

corpora used, etc.) across a range of user studies; as a comparison, the final row showcases our own study.

2.1 Interface level

Morris et al. (2008) proposed to raise awareness among collaborators about their querying actions by incorporating visual hints (such as underlying collaborators' query terms) at the search snippet level. In an earlier user study, Morris and Horvitz (2007) had shown that a shared query history among collaborators is a significant source of search engine result (SERP) views, with more than 30% of SERPs in the study being retrieved from query history clicks. Instead of just a simple shared query history, the `CoSense` system (Paul and Morris 2009) provides users with a detailed shared timeline (providing information on clicks, views and chat messages) as well as a shared workspace. Capra et al. (2012) included filtering options in the SERP, enabling collaborators to view the documents rated as (non-)relevant by their collaborators. This though did not lead to higher recall levels as often collaborators rated the same documents instead of exploring new areas of the search space. Diriye and Golovchinsky (2012) incorporated a search result histogram in the UI of their collaborative search system, enabling users to keep track of the queries that resulted in a document being retrieved.

Facilitating the division of labour—beyond providing a chat widget as offered in a number of systems, e.g. (Morris and Horvitz 2007)—has been explored for example in the `CoSearch` system (Amershi and Morris 2008) which assumes a shared-computer collaborative search setting with one main device (e.g. a Desktop) and several smaller devices (e.g. mobile phones) to enable distributed control of the search. The `SearchTogether` system (Morris and Horvitz 2007) includes a "recommendation queue" interface feature, enabling users to recommend documents to their collaborators for reading.

2.2 Techniques level

Beyond the interface level, `SearchTogether` (Morris and Horvitz 2007) also offers a "split searching" mechanism to distribute the labour (and thus avoid redundancy) with just one of the collaborators submitting a query and the search system splitting the search results in a round-robin fashion across all group members for evaluation. A more intelligent form of splitting was later proposed by Morris et al. (2008): here, each collaborator's personal profile was taken into account in the splitting process. Joho et al. (2009) approached split searching by topical clustering with every collaborator receiving the documents associated with a particular aspect of the topic but did not observe an increased search effectiveness compared to simple round-robin splitting.

A common search task type in collaborative search studies is the recall-oriented type where shared relevance feedback (query expansion based on the feedback provided by all collaborators) has been shown to increase search effectiveness compared to independent relevance feedback (query expansion based on each collaborator's feedback individually) (Joho et al. 2009).

2.3 Algorithms level

While considerable work on the techniques level investigates how best to split up the document space, on the algorithms level we consider changes made to the retrieval algorithms

themselves, in particular changes based on asymmetric *user roles*. Here, collaborating searchers are no longer treated as equals, but are either assigned fixed (Pickens et al. 2008; Shah et al. 2010; Soulier et al. 2014b) or dynamic roles (Soulier et al. 2014a) based on their search strategies and behaviours.

Pickens et al. (2008) were one of the first to propose *algorithmic mediation*, formulating two roles in a collaborative search process, each with their own specific ranking algorithm and user interface designed for their respective task: the *prospector* issues diverse queries in order to explore the search space while the *miner* acts as assessor of documents, in particular those occurring highly ranked in many of the prospector's result lists. Two alternative role types that no longer differ in their task (issuing queries vs. assessing documents) but in the type of information received were introduced by Shah et al. (2010): the *gatherer* receives result lists optimized for effectiveness while the *surveyor* receives result lists optimized for diversity. In both studies, role-based algorithmic mediation led to a higher search effectiveness than the naive merging of search results by independent pairs of searchers.

Soulier et al. later showed similar positive results when assigning collaboration roles (1) according to domain expertise (Soulier et al. 2014b) or (2) dynamically based on users' search behaviours (Soulier et al. 2014a).

Recently, Böhm et al. (2016) have developed a first formal cost model for collaborative result ranking with the aim of deriving (theoretically) optimal collaboration strategies.

2.4 Tooling

In contrast to single-user search where the research community has access to a number of well-functioning open-source search systems (such as `Indri`, `Terrier` or `Elastic`), collaborative search is hampered by a lack of open-source tooling. Although a number of collaborative search systems have been proposed (Morris and Horvitz 2007; Amershi and Morris 2008; Paul and Morris 2009; Diriye and Golovchinsky 2012; Capra et al. 2012; González-Ibáñez and Shah 2011; Bailey et al. 2012), few were made publicly accessible and only `Coagmento` (González-Ibáñez and Shah 2011) and `SearchX` (Putra et al. 2018) are being actively maintained today. Note that, in contrast to `SearchX`, `Coagmento` requires its users to install a mobile app or browser plugin, making it less well suited for crowd-sourcing experiments.

2.5 Group Size Dynamics

Let us consider once more the work of Joho et al. (2009) who investigated the impact of changing group sizes on retrieval effectiveness in a simulation study. Specifically, the query and assessment actions of collaborating groups (one to five users per group) were simulated[1] and eight different search strategies aimed at knowledge sharing and division of labour were evaluated (including independent searching, searching with judged documents removed from the SERP, query expansion with independent/shared relevance feedback, etc.). Increasing group sizes led to increased search effectiveness measured in recall, though with diminishing returns—the largest change in retrieval effectiveness ($\approx$ 50% increase) was observed when a second member entered the team, the smallest when adding

[1] Note, that the simulation made use of the Robust track 2005 data; we use the same topics and corpus in our experiments.

the fifth member (5–12% increase depending on the search strategy). Due to the simulated nature of the study, it is unclear whether those findings will also hold when real users are collaborating—we investigate this very research gap in this work.

3 Hypotheses

Before describing the study design, we now list the specific hypotheses we investigate in this work, based on the simulation findings in Joho et al. (2009)—all of them related to the impact of group size on search behaviour, effectiveness and processes:

H1 Group recall increases with increasing group size, with diminishing gains.

H2 For topics with a higher number of relevant documents, increased group size will have a relatively higher impact on group recall (as it takes more work to find all relevant documents).

H3 A large group size is more useful early in the search session, with improvements in recall over lower group sizes decreasing as the search session progresses.

H4 Division of labour and sharing of knowledge across a group of users increases their group recall; the effect is consistent across group sizes.

4 Study design

In order to investigate our research question, we extended `SearchX` as needed. As corpus we chose Aquaint. Although it is an older and rather small corpus, it is still a preferred choice for interactive IR studies as seen in Table 1 due to its clean nature (newswire texts). More recent web corpora such as ClueWeb require extensive pre-processing (spam filtering, boilerplate code removal, rendering, etc.) with unclear benefits. We selected a number of difficult Aquaint topics and ran a crowd-sourced study where groups of between one (i.e. single-user search) and six users worked collaboratively on three search tasks. To study the impact of different collaboration features we deployed three variants of our search system.

4.1 Search variants

The three search system variants we explore are listed in Table 2; we also point out their correspondence to the variants discussed by Joho et al. (2009). Variant **S-Single** does not contain any collaborative search affordance, each searcher receives a single-user search instance. In contrast, variant **S-UI-Coll** provides two interface-based awareness features: a shared query and a shared saved-documents widget making all queries and saved documents available to all collaborators (cf. Fig. 1). In addition, we also implemented a soft *division of labour* (DoL): results saved by any collaborator are by default hidden from the result listings of all collaborators in the group (though it is possible to "unhide" them); documents explicitly marked as "exclude" by any collaborator are also hidden by default. The third variant, **S-UIAlg-Coll** has the same features as **S-UI-Coll** as well as algorithmic mediation for sharing of knowledge. Concretely, we implemented *shared relevance*

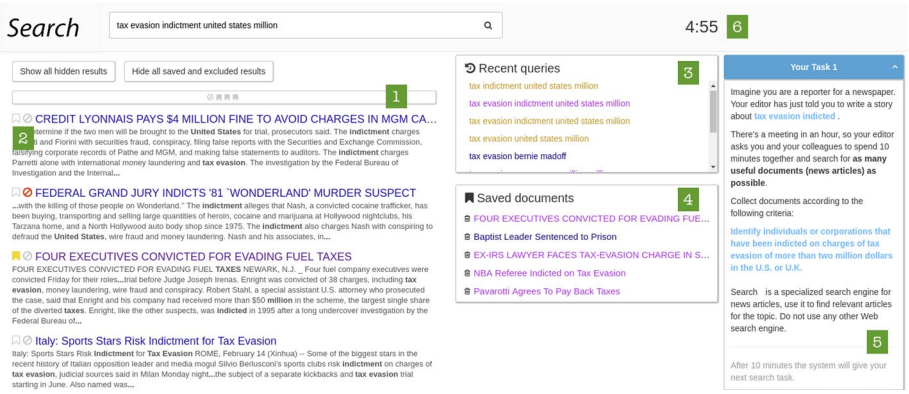


Fig. 1 Collaborative search interface: (1) shows how many results are currently hidden from the SERP because they were saved or excluded (here: 3 saved, 1 excluded) by a group member; (2) refers to the two icons to save a document or to exclude it; (3) shows the recent queries of all collaborators with colour coding to distinguish the different group members; (4) shows the saved documents, applying the same colour coding as in (3); (5) shows the task description (visible at all times) and (6) shows the timer. A click on any of the title snippets or saved documents will open the document viewer, a click on one of the recent queries will execute the query

Table 2 Overview of our collaborative search conditions and their correspondence to the variants explored by Joho et al. (2009)

S-Single	Independent search with individual bookmarks and individual query history (no awareness, no division of labour)
	Similar to *SS1* of Joho et al. (2009)'s variant "Team members performs search independently"
S-UI-Coll	**S-Single** + Shared saved documents, shared query history and collapsing of saved and excluded documents in the SERP (awareness, interface-level division of labour)
	Corresponds to *SS2* of Joho et al. (2009)'s variant "SS1 with unjudged documents only"
S-UIAlg-Coll	**S-UI-Coll** + Shared relevance feedback (awareness, interface-level division of labour and system-level sharing of knowledge)
	Corresponds to *SS4* of Joho et al. (2009)'s variant "SS3 with shared relevance feedback"

Note that across our conditions the judged documents are automatically hidden (though "unhiding" them is possible too)

feedback (RF), where the documents saved by all collaborators are employed in the query expansion stage of all collaborators.

We selected **S-Single** and **S-UI-Coll** in order to determine the benefit of adding more collaborators in a basic collaborative search setting—instead of designing an interface with as many collaborative search features as possible, we aimed at a search interface that looks familiar to today's web searchers while still providing awareness and soft division of labour features. We chose not to include a chat widget, as with increasing group sizes the existence of a chat is likely to lead to a long start-up time (users communicating and managing their searches). We chose **S-UIAlg-Coll**, as shared RF has been shown to significantly outperform all other variants in Joho et al. (2009)'s simulation study.

In addition, we explore here to what extent the simulation results hold in an experiment with actual users (i.e., crowd-workers). While crowd-workers are only an approximation of

"real" users, they are as close as we can get in a large-scale user study. We hypothesise that with users the benefits of shared RF may be outweighed by the cognitive load experienced by users whose search results no longer match their expectations (and this problem is likely to get worse as group sizes increase).

Due to the large number of study participants already required for those three search variants, we opted to not include a fourth search variant with individual relevance feedback.

4.2 System implementation

The first version of our open-source `SearchX`[2] framework was released in early 2018 (Putra et al. 2018). We here describe the extensions we added to our framework in order to investigate the research question of this work. `SearchX` is written in `node.js` and relies on the Indri IR toolkit as search back-end. The front-end with all collaborative search features enabled is shown in Fig. 1; as noted before, it does not require a user-side installation, it runs in any modern browser. As we employ the system with crowd-workers, we implemented an interactive step-by-step user interface guide to ensure that all workers are aware of all search system features and added compliance warnings (e.g., we show an alert message when a user tries to close the tab to warn them that closing the task will quit the experiment).

The two challenging issues in implementing the system were (1) the synchronising of crowd-workers and (2) the synchronisation of interface and backend-level shared division of labour; we now provide more details on these two steps in turn.

4.2.1 Synchronising crowd-workers

In lab studies, experimenters often sign up *groups* of collaborating users, e.g. (Morris and Horvitz 2007; Morris et al. 2008; Paul and Morris 2009; Kelly and Payne 2014; Joho et al. 2008; Tamine and Soulier 2015), instead of individuals that are grouped together on the fly—those groups are stable, it is unlikely that a member drops out in the middle of the experiment. In a crowd-sourcing setup, we can neither sign up groups of workers that know each other nor ensure that every worker completes the task. To overcome these issues we implemented a virtual "waiting room" where crowd-workers who signed up for our task were asked to wait up to 10 min (we also offered a game of snake to pass the time). Once the desired number of crowd-workers had signed up, or the 10 min were up, they received a shared collaborative search session (randomly assigned to one of the three search variants) and the waiting room became available for the next set of crowd-workers. In the case where the desired number of collaborators was not reached, our system split up the active collaborators into one or more different collaborative search instances, depending on the number of participating groups required for each group size. Concretely, we initially settled on evaluating collaborations between groups of sizes 1 (i.e. single-user search), 2, 4 and 6 collaborators. If for example after 10 min of waiting five workers had joined our virtual waiting room, the system created two groups (one of size 1 and one of size 4).

Lastly, since each group had to tackle three search topics, we could not rely on workers to individually move to the next search topic. We provided workers with a visible timer and

[2] http://felipemoraes.github.io/searchx.

Table 3 Selected ROBUST05 topics, including the number of relevant documents in Aquaint and the av. AP of the three top performing ROBUST05 runs

ID	Topic	#Rel.	av. AP	#Rel. cleaned
341	Airport security	37	0.08	32
367	Piracy	95	0.09	81
650	Tax evasion indicted	32	0.09	24

The last column shows the number of relevant documents after corpus cleaning

the interface switched automatically to the next topic after 10 min, ensuring that collaborators remained synchronised in their search topics.

4.2.2 Synchronised algorithmic mediation

Shared relevance feedback (utilised in **S-UIAlg-Coll**) and division of labour (utilised in **S-UI-Coll** and **S-UIAlg-Coll**) can be implemented in one of two ways: either immediately or delayed. In the immediate version, as soon as a collaborator saves a document that action should be reflected in the SERPs of all collaborators—not just by updating the shared saved documents list and hiding the document in question from the SERPs, but also by rerunning each user's submitted query with the new set of relevant documents. This is likely to confuse users as they cannot anticipate when (and why) a result ranking suddenly changes; worse still, if a user paginates through the result list, she might miss the newly highly ranked relevant documents because she is looking at lower ranks. We overcome this issue by opting for delayed RF and division of labor: only when a collaborator issues a new query are the updated saved documents included in the ranking model, and are saved and excluded results hidden. Documents that are promoted by RF to a previous page are always shown on the current page, to prevent users from missing potentially relevant documents. Note, that in this delayed update model, the status change of the saved documents widget and save/exclude buttons attached to each search result are still immediately occurring.

4.3 Corpus, Topics and Retrieval Models

We use Aquaint as our document collection; it contains 1,033,461 news articles and has been used in a number of prior studies (cf. Table 1). It was also the collection of choice in the TREC 2005 Robust track (Voorhees 2006); and we thus refer to it (Aquaint plus TREC 2005 Robust track topics) as ROBUST05. It is focused on 50 poorly performing topics in an ad hoc retrieval setting. In order to select the topics for our study, we took the best three automatic runs submitted to ROBUST05 and for each topic computed the mean of the average precision across those runs. We ranked the topics in ascending order and considered only the first ten, i.e. the topics most difficult for the best performing retrieval systems at the time. We chose those topics, as collaborative search is most appropriate for difficult search topics; note that this choice is in contrast to prior Aquaint-based collaborative search studies (cf. Table 1) that often opted for *interesting* topics, largely ignoring topic difficulty. As some of those ten topics share relevant documents, we manually selected three very different topics with the additional constraint of at least 30 relevant documents and at most 100 relevant documents in the corpus—too many or too few relevant documents will limit the insights we can gain from our study. The final three topics are listed in Table 3. We

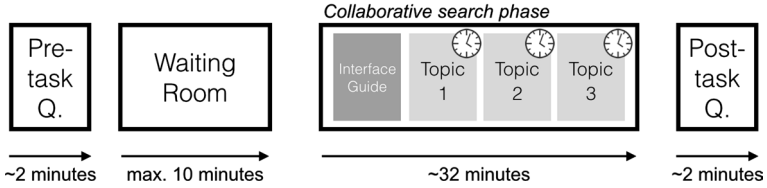


Fig. 2 Overview of a worker's flow through our system

note, that only topic 367 overlaps with the topics employed by Joho et al. (2009). According to our selection criteria, the remaining topics employed in the simulation study were not of sufficient difficulty.

We indexed Aquaint with stopword removal and Krovetz stemming using `Indri` and removed near duplicate documents as well as documents without a title (to ensure that all search result snippets look similar instead of having some with *Untitled* as title) before indexing. For near duplicate detection, we used SimHash with parameters *blocks* = 4 and *distance* = 3, following Manku et al. (2007). We also processed the relevance judgments file accordingly, ignoring all documents that we removed in our pre-processing step before computing our retrieval metrics (cf. last column of Table 3). We also relied on `Indri` to generate the snippet text in a query-dependent manner for the SERP. Overall, after the cleaning steps 854, 130 documents remained in our index.

In variants **S-Single** and **S-UI-Coll**, the retrieval algorithm is language modeling (LM) with Dirichlet smoothing (Zhai and Lafferty 2004) with hyper-parameter setting $\mu = 2500$. The relevance feedback (RF) variant **S-UIAlg-Coll** implements relevance-based language modelling (Lavrenko and Croft 2001), in particular RM2 with 10 feedback terms and all documents saved by the collaborators. In an offline experiment we found RM2 with *true* RF (as found in the official TREC relevance judgement file) to outperform the LM baseline on average by 82.65% across our three topics: we sampled 5 relevant documents per topic from the official qrels 20 times, retrieved ranked lists of results based on RM2 and LM, removed those 5 relevant documents from the result lists (and the relevance judgement file) and computed the recall. Thus, as long as our participants save (mostly) relevant documents, RF will improve the quality of the search results compared to the non-RF based language modeling variant employed in **S-Single** and **S-UI-Coll**.

4.4 Crowd-sourced task setup

We opted for a crowd-sourcing setup due to the number of participants we require: three search variants, each evaluated with ideally four group sizes (1/2/4/6). We aim for ten groups in each setup—a common size, cf. Table 1—thus requiring more than 300 participants, considerably more than in any of the listed lab studies.

We recruited workers from the Prolific platform,[3] which has been shown to be a more reliable source of workers for cognitively demanding tasks than MTurk or Figure Eight (formerly CrowdFlower), two other popular crowd-sourcing platforms (Peer et al. 2017). Workers can only participate once in our study. Each worker, once accepting the task, is

[3] https://www.prolific.ac/.

> Imagine you are a reporter for a newspaper. Your editor has just told you to write a story about [**ROBUST05 topic title**]. There's a meeting in an hour, so your editor asks you and your colleagues to spend 10 minutes together and search for as many useful documents (news articles) as possible and save them. Collect documents according to the following criteria: [**ROBUST05 topic description**].

Fig. 3 Task template

directed to our server and moves through the workflow depicted in Fig. 2: the pre-task questionnaire contains a description of collaborative search and four collaborative search questions borrowed from Morris (2013) to prime the workers for the upcoming collaborative search tasks;[4] the waiting room component contains an explanation of the "waiting room" concept, a visible timer, the option to play a game of snake and the option of an audio alert, to enable the worker to use other browser tabs while waiting for sufficiently many workers to join. Once workers move to the collaborative search phase, they first receive an interactive tour of the search interface before starting to work on their assigned topics. To provide context for the ad hoc search topics, we employed the task template in Fig. 3, inspired by previous studies (Azzopardi et al. 2013; Kules and Shneiderman 2008). After 10 min of searching, all collaborators are automatically moved to the next search task to ensure synchronisation. This recall-oriented task can be found in settings such patent-retrieval and e-discovery and represents a typical task in collaborative search which induces a complex and exploratory behaviour from users as discussed by Morris (2013).

The post-task questionnaire contains the following seven questions on search satisfaction:

1. How many people did you just now collaborate with (not including yourself)? *[Number]*
2. The color coding of the query history and bookmarks made sense to me. *5-Likert scale [Disagree, Agree]*
3. It was easy to understand why documents were retrieved in response to my queries. *5-Likert scale [Disagree, Agree]*
4. I didn't notice any inconsistencies when I used the system. *5-Likert scale [Disagree, Agree]*
5. It was easy to determine if a document was relevant to a task. [Disagree, Agree]
6. How difficult was this task? *5-Likert scale [Very easy, Very difficult]*
7. Did you find the collaborative features useful (multi-grid question with one row for each feature: recent queries, saved documents, and hiding saved and excluded results)? *5-Likert scale [Disagree, Agree]*

Workers are assigned to search variants and group sizes at random; the order of the three search topics is randomised per group.

[4] e.g., *Think about the most recent time you collaborated with others to search the web. Describe what were you looking for.*

Table 4 Number of collaborating groups across search variants, topics and group sizes

	Topic ID	{1}	{2}	{3,4}	{5,6}
S-Single	650	12	–	–	–
	367	12	–	–	–
	341	12	–	–	–
S-UI-Coll	650	11	12	10	9
	367	12	11	10	9
	341	13	10	11	8
S-UIAlg-Coll	650	17	8	16	12
	367	17	11	13	13
	341	19	10	14	13

For **S-Single** we simulate the collaborative search behaviour across larger group sizes with the data collected from the single-user search data

4.5 Post-processing of collected logs

Due to the unpredictable nature of crowd-workers, a synchronised collaborative search experiment is not easy to conduct. As group sizes increase, it becomes more difficult to form groups (as sufficiently many crowd-workers have to choose the task at roughly the same time) and worker dropout becomes more likely during the task. We mitigate these issues in three steps:

1. On a *per topic* basis, we only consider collaborators as *active* in a group that issued at least one query for the topic.
2. We consider the number of collaborators in a group on a topic-by-topic basis (we only count active collaborators), instead of fixed across all three search topics. For instance, if a group starts with four workers, and one worker becomes inactive after the first search topic and another worker drops at the start of the third topic, we consider this as a group of four collaborators for the first topic, a group of three for the second topic and a group of two for the third topic.
3. As after these two steps we have groups of three and five collaborators, we subsume the logs of groups of 3–4 and 5–6 collaborators respectively into two groups in the analyses that follow. Overall, we thus analyse four group sizes: $\{1, 2, \{3, 4\}, \{5, 6\}\}$. We decided on this merging strategy (instead of for instance merging together groups of 2–3 participants) as Joho et al. (2009) found in their simulation study the addition of the second collaborator to bring about the greatest benefit in terms of search effectiveness.

Table 4 provides an overview of how many groups in total (67 groups with between 2 and 6 participants—recall that a group has three topics to participate in) participated in our experiment across all search variants after the above post-processing steps. As **S-Single** is the single-user search setup, we collect 12 instances of single-search tasks and then *simulate* the behaviour across larger group sizes by grouping users together and merging their saved documents in line with prior works (Pickens et al. 2008; Joho et al. 2009; Htun et al. 2015; Foley and Smeaton 2010). We group users together by considering all possible combinations for each group size, ensuring that the data for each user is weighed equally in the results.

A total of 335 workers participated in our study, of which 30 were excluded since they did not perform any actions. We thus had 305 valid participants. In Table 4 we have the largest number of groups (16) for condition **S-UIAlg-Coll** and groups $\{3, 4\}$. This artefact can be explained by the fact that regularly participants assigned to collaborating in a team of six (which required the longest waiting time) grew impatient and dropped out, and thus often the result were formed groups of three or four participants. On average, our workers spent 42 min on the task, including the at most 10 min in the virtual waiting room. We paid £3.75 for the task.

The drop-out rate of participants *during* the experiment was 30.4%, that is the rate of groups starting off with at least two collaborators that decreased in size whilst working through the three topics.

4.6 Evaluation metrics and tests

We use the following metrics and statistical analyses to compare the search variants and group sizes in line with previous works by Joho et al. (2008) and Pickens et al. (2008).

Retrieval effectiveness in order to measure retrieval effectiveness, we employ group recall which is defined as the recall for the union of the sets of documents that each collaborator saved. This metric is appropriate, given the fact that we designed our search tasks to be recall-oriented. The group recall $GR(g, t)$ is calculated for each group g and topic t. To calculate the average group recall $AGR(s, t)$ for all groups for size s and topic t, group recall is averaged across all groups in a given size category:

$$AGR(s,t) = \frac{1}{|G_{s,t}|} \sum_{g \in G_{s,t}} GR(g,t). \tag{1}$$

Temporal analyses at a number of time points during a search session—specifically, after $\{2, 4, 6, 8, 10\}$ min—the group recall for each group is computed; here, we only take documents saved until that point in time into account. The search session start time is fixed to the time the first member of a collaborating group submits a query. Average group recall for different group sizes is computed in the same manner as discussed above.

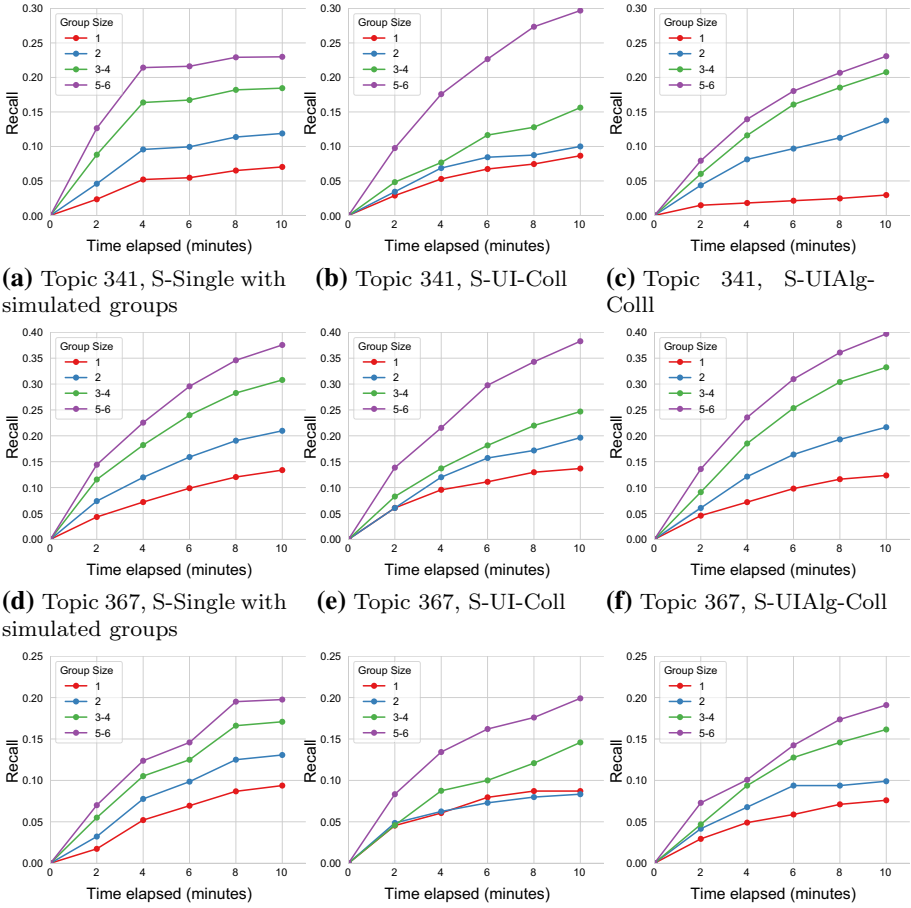
Statistical analyses in order to compare the impact of group size and search variant in Sect. 5, we conducted a two-factor analysis of variance (ANOVA) separately for each topic. We examined the ANOVA assumptions with Levene's test (homogeneity of variances) and Shapiro-Wilk's test (normality of the ANOVA residuals). We conducted a *post-hoc* analysis using Kruskal-Wallis for the **S-Single** search variant and Tukey's HSD test for the other search variants.

5 Results

We now first present an overview of the main outcomes of our study and provide insights into our participants' search behaviours. We then discuss these results in light of the research hypotheses listed in Sect. 3.

The main results of our study are shown in Fig. 4 where we present a detailed overview of the development of average group recall across time, different group sizes and search variants. A number of observations can be made:

- in all cases a larger group size leads to a larger recall level;



(a) Topic 341, S-Single with simulated groups **(b)** Topic 341, S-UI-Coll **(c)** Topic 341, S-UIAlg-Colll

(d) Topic 367, S-Single with simulated groups **(e)** Topic 367, S-UI-Coll **(f)** Topic 367, S-UIAlg-Coll

(g) Topic 650, S-Single with simulated groups **(h)** Topic 650, S-UI-Coll **(i)** Topic 650, S-UIAlg-Coll

Fig. 4 Overview of the average group recall for each topic and search variant computed in 2-min time intervals

- in line with Joho et al. (2008)—and in contrast to Shah and González-Ibáñez (2011)—we did not observe a synergy effect: pairs of collaborators were *not* more effective than two independent searchers whose results were aggregated.

Table 5 lists the average group recall on a per-topic basis, with statistical differences in recall across group sizes and search variants highlighted—note that due to the nature of **S-Single** (where group sizes > 1 are simulated) it is not possible to reliably test for statistical differences between **S-Single** and **S-UI-Coll/S-UIAlg-Coll**. We find that group sizes of $\{3, 4\}$ and $\{5, 6\}$ respectively to lead to significantly higher recall levels than smaller groups.

To provide insights into participants' search behaviours as well as to ascertain that our participants conducted searches as intended, we list major characteristics of their individual behaviour in Table 6. In order to aggregate participants' behavioural traces in a meaningful

Table 5 Topic-wise average group recall (averaged across all groups in a single topic/search-variant) across conditions and group sizes

	Group size	Average group recall per topic		
		650	367	341
S-Single	1	0.094	0.134	0.070
	2	0.131	0.210	0.119
	3–4	0.171	0.308	0.184
	5–6	0.198	0.376	0.230
S-UI-Coll	1	0.087	0.137	0.087
	2	0.083	0.196	0.103
	3–4	0.146	0.249	0.159^{A1}
	5-6	$0.208^{U1,U2,A1}$	$0.391^{U1,U2,A1}$	$0.305^{U1,U2,U34,A1,A2}$
S-UIAlg-Coll	1	0.076	0.125	0.031
	2	0.109	0.231	0.138
	3–4	$0.169^{U2,A1}$	$0.349^{U1,U2,A1}$	$0.214^{U1,A1}$
	5-6	$0.219^{U1,U2,A1,A2}$	$0.404^{U1,U2,U34,A1,A2}$	$0.243^{U1,U2,A1}$

Statistical significance was determined via Tukey's HSD test independently for each topic; in each topic column, significant improvements at $p < 0.01$ are marked with superscript ^{XY} where X is the variant ('U' in the case of **S-UI-Coll** and 'A' in the case of **S-UIAlg-Coll**) and Y is the respective group size. For the **S-Single** simulated groups we determined significant values among group sizes only within **S-Single** via Kruskal-Wallis test independently for each topic (we omitted superscript symbols as all group sizes shows significant different results at $p < 0.01$)

Table 6 Overview of individual search behaviours across the search conditions

	GS	#Q.	Query length (#words)	#Viewed docs.	Viewing doc. time (#s)	#Unique saved docs.
S-Single	1	5.50	3.72	11.83	12.31	10.67
S-UI-Coll	1	7.50	3.58	8.00	12.55	10.67
	2	8.67	3.57	5.67	10.52	6.50
	3–4	8.67	3.71	7.50	9.80	7.50
	5–6	8.67	4.26	6.33	8.89	7.17
S-UIAlg-Coll	1	6.67	3.79	12.50	11.49	8.50
	2	6.33	3.34	10.67	8.62	9.67
	3–4	6.33	4.06	7.00	9.51	8.17
	5–6	8.33	3.99	7.33	8.64	7.00

GS is the group size and #Q. the number of queries. For each topic, the median value is computed; reported here is the average of the those three median values

manner, we resorted to computing the median value across all participants of a certain collaboration group size, topic and search variant. Since we have three topics in total, we then computed the average of the three median values. We observe across search variants and group sizes, that the median number of queries issued by a searcher for a topic varies between 5 and 9 with queries being of moderate length (3–4 terms). We find a decreasing trend in the number of viewed documents as well as the amount of time spent on each

document viewed for participants in larger collaborating groups—there are two possible explanations for this trend: on the one hand, the participants may be more occupied with the activities of their collaborators (as they appear in the shared query history and the shared bookmarking widgets) or, on the other hand, the participants may be more complacent, knowing that their collaborators are active in the same search task. The last column in Table 6 shows that complacency is not a likely explanation, as the number of unique saved documents remains relatively stable across collaborative group sizes (with the exception of the single-user case). Finally, we note that the document viewing time is rather short (between 8 and 13 seconds in Table 6); this can be explained by the fact that the Aquaint documents themselves are relatively short, with an average length of 438 words.

Based on these statistics and a manual check of a sample of search logs our participants generated we conclude that our participants provide valid log traces for our analyses.

5.1 Search effectiveness across group sizes

The first observation we make is with respect to recall: for none of the topics, search variants and group size combination is the reported recall greater than 0.4, indicating the difficulty of the topics and the potential benefit an increasing collaborator pool could bring about. We also see that, despite picking the most difficult topics, the maximum recall varies considerably (Fig. 4), with a maximum of 0.2 (topic 650), 0.3 (topic 341) and 0.4 (topic 367) respectively.

While the trends across topics and search variants in Fig. 4 are similar, there are two apparent "outliers": (b) shows a significant recall gap (the recall doubles) between group sizes of $\{3,4\}$ and $\{5,6\}$ and (h) shows no change in recall between group sizes of one and two. We did not observe anomalies in the search logs across those two topic/variant setups; based on this finding we argue that these are slight variations are a result of our user study setup.

Across all search variants and topics we find the first part of **H1** (*Group recall increases with increasing group size ...*) to be supported. In **S-Single** each additional group member results in a similar absolute increase in recall levels at the end of the search session, i.e. at the 10 min mark in Fig. 4.

In contrast, for the interface-based collaboration variant **S-UI-Coll** we find only small differences in recall level between teams of one (i.e. a single searcher) and teams of two collaborators at the end of the search session. The largest increase in recall (ranging from + 42 to + 92% depending on the topic) occurs when moving from groups of $\{3,4\}$ to groups of $\{5,6\}$ collaborators. The significance tests reported in Table 5 show that within **S-UI-Coll** for all topics the largest group size yields significantly better recall levels than group sizes of 1 and 2. In case of topic 341 the difference is also significant with respect to group size $\{3,4\}$ for the **S-UI-Coll** variant.

A somewhat different picture once more emerges when considering algorithmic mediation (**S-UIAlg-Coll**): here, we find smaller differences in recall (between + 14 and + 30% depending on the topic) when moving from $\{3,4\}$ collaborators to $\{5,6\}$ compared to moving from two to $\{3,4\}$ collaborators (+ 51 to + 55%). These findings show that the second part of **H1** (... *with diminishing gains*) does not hold when moving from a pure simulation study to a user experiment. In line with our findings for **S-UI-Coll** we here also observe statistically significant differences in recall between the largest group size and group sizes 1 and 2. We also note that in contrast to Joho et al. (2009) we do not observe a convergence of the recall levels across group sizes towards the end of the search session—that

is to say, while we evaluated no more than six collaborators (as crowd-sourcing becomes more difficult with increasingly high group sizes), based on our results we can expect even larger group sizes to yield additional increases in recall. One explanation for this difference can be found in the fact that we focused specifically on difficult topics, while in Joho et al. (2009)'s simulation study easy topics are included which reach high recall levels even across a small number of collaborators.

We now move on to hypothesis **H2**, which states: *For topics with a higher number of relevant documents, increased group size will have a relatively higher impact on group recall*. Topic 367 has 95 relevant documents, more than double that of topics 341 and 650 with 37 and 32 relevant documents respectively (cf. also Table 3). When we consider the recall developments in Fig. 4 across the different topics, we have to find **H2** to not hold: for topic 367 we do not observe a higher impact on group recall than for the other two topics. Our results indicate that the choice of collaboration (interface-based only vs. algorithmic) is a more important factor with respect to explaining the impact of group size on group recall level changes.

5.2 Search effectiveness across time

Hypothesis **H3** is concerned with the development of recall over time, with simulation studies indicating that a large group size is beneficial in particular early on in the search process. We restate it here: *A large group size is more useful early in the search session, with improvement in recall over lower group sizes decreasing as the search session progresses*. Once again we consider the recall developments in Fig. 4. We find that in practice the benefit remains relatively consistent, i.e., it is not restricted to the early minutes of the search session. As time progresses, the smaller collaborating groups do not "catch up" with the larger groups in terms of recall; this behaviour holds across all three topics and search variants. Although we do not know what would happen after the 10 min mark (as we fixed the end of a topic's search session), we observe the recall curves for group sizes of 1 and 2 to level off somewhat across most topics and search variant combinations, in contrast to the larger group sizes.

5.3 Division of labour

Let us now consider hypothesis **H4** which states: *Division of labour and sharing of knowledge across a group of users increases their group recall; the effect is consistent across group sizes*. Recall that **S-UI-Coll** provides interface-level division of labour while **S-UIAlg-Coll** provides both interface-level division of labour and algorithmic sharing of knowledge via shared RF. In contrast to our expectations (simulations found division of labour and sharing of knowledge to increase group recall compared to post-hoc merging of result lists as done in **S-Single**) we find **S-Single** to perform on par with **S-UI-Coll** and **S-UIAlg-Coll** in terms of retrieval effectiveness. While providing sharing of knowledge (**S-UI-Coll** vs. **S-UIAlg-Coll**) does not yield significant changes in recall level for a given topic and group size, we find that for 8 out of 9 topic/group-size comparisons (ignoring groups of size one) **S-UIAlg-Coll** reports a higher recall level than **S-UI-Coll**, providing some support for **H4**.

Lastly, we determine whether our participants actually engaged with their collaborators through our interface affordances. To this end, in Table 7 we report the median number of click interactions our groups of collaborators had with queries from the `Recent`

Table 7 Usage of collaborative search interface features by groups of collaborators

	Group size	#Clicked queries	#Viewed saved docs.
S-UI-Coll	2	0.00	0.00
	3–4	0.50	0.67
	5–6	4.00	1.50
S-UIAlg-Coll	2	0.00	0.00
	3–4	0.33	0.00
	5–6	4.33	2.00

Included are only clicks on queries and saved documents by collaborators that did not issue (save) the original query (document). For each topic, the median value is computed; reported here is the average of the those three median values

Table 8 Group size vs. perceived group size in % across search variants

Condition	Group size	Perceived group size in %						
		1	2	3	4	5	6	7+
S-Single	1	*50*	42	0	8	0	0	0
S-UI-Coll	1	*90*	0	0	10	0	0	0
	2	6	*81*	13	0	0	0	0
	3–4	4	23	*31*	*15*	19	4	4
	5–6	3	11	20	43	*9*	*6*	8
S-UIAlg-Coll	1	*69*	19	0	0	6	6	0
	2	0	*44*	19	25	6	0	6
	3–4	6	26	*35*	*12*	15	6	0
	5–6	4	19	10	45	*12*	*10*	0

Results reported based on the post-questionnaire (question 1, cf. Sect. 4.4). Shown in italic is the cell value where actual=perceived group size

`Queries` and documents from the `Saved Documents` widgets. Here, we ignore interactions of collaborators with their own posed queries and their own saved documents. We find that with increasing collaboration group size more such interactions take place, though overall their number remains small (e.g. for groups of size {5,6} the median number of query widget interactions is 4–5, depending on the topic). This is in contrast to the findings by Morris and Horvitz (2007) who reported in the evaluation of their `SearchTogether` system that more than 30% of SERP views could be traced back to query history clicks. It should be noted though, that `SearchTogether` was a standalone Desktop client with very elaborate collaboration widgets, while we strove to make the search experience collaborative but still very much relatable to modern web search.

5.4 Worker perceptions

Finally, we analyse the participants' responses to the seven questions in the post-questionnaire (the questions are listed in Sect. 4.4). For our analysis, we only consider responses from participants in stable groups, i.e. groups that maintained their original size across all

three search topics (as one of the questions is concerned with the *perceived* group size). We thus report responses from 12, 31, and 43 participants in search variants **S-Single**, **S-UI-Coll** and **S-UIAlg-Coll**, respectively. The most interesting finding pertains to the perceived versus actual group size[5] as reported in Table 8: while for small group sizes (single-user search or pairs of collaborators) almost always the majority of participants is able to pick the correct group sizes, with increasing group size the perception varies widely, with the vast majority of participants underestimating the size of the group at actual group size {5–6}. Depending on the search variant, we also see a considerable number of participants in the single-user search condition to report themselves having been in a collaboration—we attribute this to the priming questions on collaborative search, the participants had received at the start of the experiment as well as the virtual waiting room time the participants experienced.

Questions two to six in our post-questionnaire focus on the participants' search experience, specifically (Q2) color coding, (Q3) easy to understand document retrieval, (Q4) no inconsistencies, (Q5) easy to determine relevance, (Q6) task difficulty. The feedback for questions Q2–Q5 is similar across the search variants and group sizes, ranging from 3–4 on the 5-point Likert scale, and thus we can conclude that the general search experience was positive. For Q6, we also find similar task difficulty values across the three search variants ranging from average values of 2.75 (**S-Single**)–2.79 (**S-UIAlg-Coll**) and 2.84 (**S-UI-Coll**).

With respect to the usefulness of the different interface features (Q7), for all the search variants, participants agreed the saved documents to be the most useful feature, followed by the recent queries and then the hiding of already saved/excluded search results.

6 Conclusions

The impact of group size on collaborative search effectiveness has not received a lot of attention in past research. In particular, we are aware of only one work that focuses on this issue: Joho et al. (2009), who performed elaborate simulations to investigate the effect of group size changes in recall-oriented search tasks. Simulations though, are limited in their ability to model the real world and thus we conducted an elaborate user study to investigate to what extent the findings of this simulation study hold in a setup with actual users.

We designed a crowd-sourcing based user experiment and extended our synchronous collaborative search framework `SearchX` capable of synchronising crowd-workers across tasks and user groups to explore the extent to which the simulation results hold in practice. Of the four hypotheses we investigated (all focusing on the impact of group size changes on search effectiveness), we find partial support for only two of them (**H1** and **H4**), demonstrating ultimately the limitations that a simulation setup suffers from.

Specifically, we do not observe diminishing returns with increasing group sizes; the group recall steadily increases as more collaborators participate in the search. We also do not find larger collaborating groups to mostly be beneficial at the start of a search session, instead, the increased recall obtained early in the search session in contrast to smaller collaborating groups is retained throughout the search session. Lastly we note, that our results also confirm our intuition that division of labour and sharing of knowledge

[5] Note that while the question in the questionnaire asked the participant for the number of collaborators *not including him/herself*, we here report the perceived group size number to simplify the comparison.

approaches—which work well in simulations as they assume perfect relevance judgement capabilities and no increased cognitive load with increasing group size—need to be considered with care as group sizes increase.

We believe that the results we presented in this work are an important step towards a greater research emphasis on changing and importantly *increasing* group sizes in collaborative search settings. In our experiments we considered a maximum of six collaborators due to inherent limitations that the mix of synchronous collaborative search and crowdworking platforms suffer from, however, at the same time we did not find evidence that we already reached the upper recall bound with our maximum group size.

Our work is limited to the remote collaborative search setting and we acknowledge that results for larger group sizes may not generalise to the co-located setting across all collaborative search dimensions. In particular, in this work we limited the communication dimension, which—in the co-located setting—is hard to control.

In future work we will address a number of limitations of our own work as well as novel avenues:

- We believe based on our experiences in this work that we have reached close to the maximum number of collaborators we can engage synchronously in a crowd-sourcing setup and as a next step will investigate the deployment of our collaborative search system in a large-scale learning environment (a MOOC) where thousands of users work together towards a shared goal. Search as learning (Collins-Thompson et al. 2017) is a relevant research area here that has recently gathered increasing attention.
- We will investigate the modelling of the collaborative search process via the recently introduced economic models of search (Azzopardi 2014) in order to gain a better theoretic understanding of the interface and algorithmic mediation approaches that are worth exploring further.
- Lastly, we will broaden our investigation towards types of search tasks that lend themselves less easily towards unambiguous evaluations such as complex exploratory search tasks on the open web.

Acknowledgements This research has been supported by NWO projects LACrOSSE (612.001.605) and SearchX (639.022.722).

References

Amershi, S., & Morris, M. R. (2008). Cosearch: A system for co-located collaborative web search. In *CHI'08* (pp. 1647–1656).

Azzopardi, L. (2014). Modelling interaction with economic models of search. In *SIGIR'14*.

Azzopardi, L., Kelly, D., & Brennan, K. (2013). How query cost affects search behavior. In *SIGIR'13* (pp. 23–32).

Bailey, P., Chen, L., Grosenick, S., Jiang, L., Li, Y., Reinholdtsen, P., Salada, C., Wang, H., & Wong, S. (2012). User task understanding: A web search engine perspective. In *NII shonan meeting on whole-session evaluation of interactive information retrieval systems, Kanagawa, Japan*.

Böhm, T., Klas, C.-P., & Hemmje, M. (2016). Towards a probabilistic model for supporting collaborative information access. *Information Retrieval Journal*, *19*(5), 487–509.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*(3), 1465–1477.

Capra, R., Chen, A. T., Hawthorne, K., Arguello, J., Shaw, L., & Marchionini, G. (2012). Design and evaluation of a system to support collaborative search. *Proceedings of the Association for Information Science and Technology*, *49*(1), 1–10.

Collins-Thompson, K., Hansen, P., & Hauff, C. (2017). Search as learning (Dagstuhl Seminar 17092). *Dagstuhl Reports, 7*(2), 135–162. ISSN 2192-5283. https://doi.org/10.4230/DagRep.7.2.135. http://drops.dagstuhl.de/opus/volltexte/2017/7357.

Diriye, A., & Golovchinsky, G. (2012). Querium: A session-based collaborative search system. In *ECIR'12* (pp. 583–584).

Foley, C., & Smeaton, A. F. (2010). Division of labour and sharing of knowledge for synchronous collaborative information retrieval. *IPM*, *46*(6), 762–772.

Gao, Y., Reddy, M., & Jansen, B. J. (2016). Shop together, search together: Collaborative e-commerce. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems* (pp. 2081–2087). ACM.

Golovchinsky, G., Pickens, J., & Back, M. (2009). A taxonomy of collaboration in online information seeking. arXiv preprint arXiv:0908.0704.

González-Ibáñez, R., & Shah, C. (2011). Coagmento: A system for supporting collaborative information seeking. *ASIST*, *48*(1), 1–4.

González-Ibáñez, R., Haseki, M., & Shah, C. (2013). Let's search together, but not too close! An analysis of communication and performance in collaborative information seeking. *IPM*, *49*(5), 1165–1179.

Hansen, P., & Järvelin, K. (2005). Collaborative information retrieval in an information-intensive domain. *IPM*, *41*(5), 1101–1119.

Htun, N. N., Halvey, M., & Baillie, L. (2015). Towards quantifying the impact of non-uniform information access in collaborative information retrieval. In *SIGIR'15*, (pp. 843–846).

Htun, N. N., Halvey, M., & Baillie, L. (2017). How can we better support users with non-uniform information access in collaborative information retrieval? In *CHIIR'17*, (pp. 235–244).

Joho, H., Hannah, D., & Jose, J. M. (2008). Comparing collaborative and independent search in a recall-oriented task. In *IIiX'08* (pp. 89–96).

Joho, H., Hannah, D., & Jose, J. M. (2009). Revisiting IR techniques for collaborative search strategies. In *ECIR'09* (pp. 66–77).

Kapetanios, E. (2008). Quo vadis computer science: From turing to personal computer, personal content and collective intelligence. *Data & Knowledge Engineering*, *67*(2), 286–292.

Kelly, R., & Payne, S. J. (2014). Collaborative web search in context: A study of tool use in everyday tasks. In *CSCW'14* (pp. 807–819).

Kules, B., & Shneiderman, B. (2008). Users can change their web search tactics: Design guidelines for categorized overviews. *Information Processing and Management*, *44*(2), 463–484.

Lavrenko, V., & Croft, W. B. (2001). Relevance based language models. In *SIGIR'01* (pp. 120–127).

Manku, G. S., Jain, A., & Das Sarma, A. (2007). Detecting near-duplicates for web crawling. In *WWW'07*, (pp. 141–150).

Morris, M. R. (2007). Collaborating alone and together: Investigating persistent and multi-user web search activities. In *SIGIR'07*, (pp. 23–27).

Morris, M. R. (2008). A survey of collaborative web search practices. In *CHI'08* (pp. 1657–1660).

Morris, M. R. (2013). Collaborative search revisited. In *CSCW'13* (pp. 1181–1192).

Morris, M. R., & Horvitz, E. (2007). Searchtogether: An interface for collaborative web search. In *UIST'07* (pp. 3–12).

Morris, M. R., Teevan, J., & Bush, S. (2008). Enhancing collaborative web search with personalization: Groupization, smart splitting, and group hit-highlighting. In *CSCW'08* (pp. 481–484).

Paul, S. A., & Morris, M. R. (2009). Cosense: Enhancing sensemaking for collaborative web search. In *CHI'09*, (pp. 1771–1780).

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., & Back, M. (2008). Algorithmic mediation for collaborative exploratory search. In *SIGIR'08*, (pp. 315–322).

Putra, S. R., Moraes, F., & Hauff, C. (2018). Searchx: Empowering collaborative search research. In *ACM SIGIR*.

Shah, C. (2010). Collaborative information seeking: A literature review. In *Advances in librarianship*, (pp. 3–33). Emerald Group Publishing Limited.

Shah, C., & González-Ibáñez, R. (2011). Evaluating the synergic effect of collaboration in information seeking. In *SIGIR'11* (pp. 913–922). ACM.

Shah, C., Pickens, J., & Golovchinsky, G. (2010). Role-based results redistribution for collaborative information retrieval. *IPM*, *46*(6), 773–781.

Smeaton, A. F., Lee, H., Foley, C., & McGivney, S. (2007). Collaborative video searching on a tabletop. *Multimedia Systems*, *12*(4–5), 375–391.

Smyth, B., Freyne, J., Coyle, M., Briggs, P., & Balfe, E. (2004). I-SPY—anonymous, community-based personalization by collaborative meta-search. In *Research and development in intelligent systems XX*, (pp. 367–380). Berlin: Springer

Soulier, L., Shah, C., & Tamine, L. (2014a). User-driven system-mediated collaborative information retrieval. In *SIGIR'14* (pp. 485–494).

Soulier, L., Tamine, L., & Bahsoun, W. (2014b). On domain expertise-based roles in collaborative information retrieval. *IPM*, *50*(5), 752–774.

Tamine, L., & Soulier, L. (2015). Understanding the impact of the role factor in collaborative information retrieval. In *CIKM'15* (pp. 43–52).

Teevan, J., Morris, M. R., & Bush, S. (2009). Discovering and using groups to improve personalized search. In *WSDM'09* (pp. 15–24).

Twidale, M. B., Nichols, D. M., & Paice, C. D. (1997). Browsing is a collaborative process. *IPM*, *33*(6), 761–783.

Voorhees, E. M. (2006). The TREC 2005 robust track. In *ACM SIGIR Forum* (Vol. 40, pp. 41–48). ACM.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *TOIS*, *22*(2), 179–214.