An AID for Avoiding Inadvertent Disclosure: Supporting Interactive Review for Privilege in E-Discovery

Jyothi K. Vinjumur
College of Information Studies
University of Maryland
College Park, MD USA
jyothikv@umd.edu

Douglas W. Oard
College of Information Studies & UMIACS
University of Maryland
College Park, MD USA
oard@umd.edu

Amittai Axelrod
UMIACS
University of Maryland
College Park, MD USA
amittai@umiacs.umd.edu

ABSTRACT

When searching for evidence in civil litigation, parties to a lawsuit have the right to withhold some content on grounds of specific privileges that serve to foster socially desirable outcomes such as open communication between attorneys and their clients. As inadvertent disclosure of privileged content can adversely impact a client's interests, review for privilege is a high-stakes process that is most often performed manually. Because the circumstances in which privilege can be claimed are generally well defined, review for privilege is amenable to some degree of automation. This paper describes the design of an interactive system to support privilege review in which the goals are to improve the speed and accuracy of privilege review. Results are reported for a within-subjects study in which six reviewers with different levels of expertise examined email for attorney-client privilege or any other valid basis for withholding the content from release. Quantitative results indicate that substantial and statistically significant improvements in recall can be achieved, but no significant differences in average review speed were detected. Participants self-reported that the identity features exposed by the system were most useful to them, and that the present implementation of features based on content or date added no discernible additional value.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Experimentation, Human Factors

Keywords

Privilege Review, E-Discovery, Email Classification

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHIIR '16, March 13-17, 2016, Carrboro, NC, USA.
Copyright 2016 ACM 978-1-4503-3751-9/16/03...$15.00.
http://dx.doi.org/10.1145/2854946.2854964.

1. INTRODUCTION

In 1989, a court in Washington, DC granted a temporary restraining order to preserve a collection of electronic messages that had been shared between members of the National Security Council in the Executive Office of the President of the United States [1]. The basis for this order was a claim that electronic messages can constitute records of the activity of an organization. In 2006, the Federal Rules of Civil Procedure were amended to make it clear that all forms of electronically stored information, including email, were within the scope of evidence that could be requested from a counterparty incident to civil litigation in the United States. Thus was born the multi-billion dollar industry that has come to be called "e-discovery."[1] The high cost of e-discovery results from two factors: (1) because the standard for relevance is expansive, large numbers of relevant documents could be found, and (2) parties can assert privilege on some relevant documents to withhold some content. Hence each of the relevant documents must typically be reviewed for privilege.

The cost of review for relevance can be controlled using text classification techniques (e.g., based on supervised machine learning), but attorneys are naturally reluctant to trust fully automated techniques for privilege review so long as the scale of the privilege review (i.e., the number of relevant documents) is not so great as to preclude manual review. For this reason, our work in this paper is focused on enhancing the performance of human reviewers. Several types of privilege might be asserted, but in this paper we focus principally on attorney-client privilege. The rationale behind attorney-client privilege is that justice will be best served when attorneys can communicate freely with their clients (e.g., on matters of fact, intent, or legal strategy), and open communication can be fostered by prospectively protecting such communication from disclosure.

Our basic approach to supporting privilege review is to train automated annotators[2] to label specific components of a message with information that we expect might help a reviewer to rapidly make a correct decision. We use a total of five annotators to enrich three types of components: people (or, more specifically, the email addresses for senders and recipients of a message), terms (words found in the message

[1]"Discovery" in this context refers to a stage in civil litigation in which parties are entitled to request from each other evidence that they believe may be relevant to the case.
[2]We use the word "annotator" to refer to an automated system that generates the assistance.

or in attachments to the message), and the date (on which the message was sent). In each case, we compute a numerical score for which higher values indicate a greater likelihood of privilege [18]; for people we also annotate job responsibilities (when known) or organization type (when known, if the job responsibilities are not known).

We have studied the usefulness of these features to human reviewers using a within-subjects user study in which six lawyers each reviewed two sets of documents (email messages, together with their attachments), one set using a baseline system with no annotations, and the second set using our AID system (named for our goal of Avoiding Inadvertent Disclosures) in which annotations were shown for people, terms, and dates. Quantitative measures of review accuracy (e.g., precision and recall) and of review speed are augmented with analysis of self-reported response to questionnaires and interviews. We seek to answer three research questions:

- Do the accuracy of the user's privilege review judgments improve when system-generated annotations are presented during privilege review?
- Does the user's review speed improve when system-generated annotations are presented during privilege review?
- Which system-generated annotations do users believe are most helpful?

Our results indicate that recall can be enhanced by displaying annotations. Although the improvements in recall come at some cost in precision, given the nature of this application, that cost may be acceptable. Participants in the study principally attribute the beneficial effects to annotations of people (rather than of terms or of dates). These formative evaluation results have implications for annotator and interface design.

The remainder of this paper is organized as follows. In Section 2 we begin by reviewing the nature of attorney-client privilege and then we introduce the document collection that we used in our study. Section 3 describes the design of our AID system. We follow this with a description of our user study design in Section 4. Section 5 presents and discusses our results, and in Section 6 we conclude with some remarks on future work.

2. BACKGROUND

2.1 Privilege Review

Privilege in legal context is a right given to the parties in a lawsuit to provide protection against the involuntary disclosure of information. Attorney-client privilege in particular exists to protect the information exchange between "privileged persons" for the purpose of obtaining legal advice. Privileged persons include [8]: (1) the client (an individual or an organization), (2) the client's attorney, (3) communicating representatives of either the client or the attorney, and (4) other representatives of the attorney who may assist the attorney in providing legal advice to the client. However, privilege does not arise simply because privileged persons communicate; it can only be claimed when the content of the communication merits the claim. For example, an email from Jeff Skilling (Enron's president) sent only to James Derick (Enron's general counsel) about pending litigation would be privileged; an email with the same content sent to both James Derrick and a personal friend of Skilling's who was not involved in Enron's business operations would not be, and an email from James Derrick to Skilling that indicated (only) his intent to resign in order to spend more time with his family also would not be privileged.

In e-discovery, documents that are initially marked as responsive to a production request (i.e., a specific request for evidence by the counterparty) are then typically subjected to a linear manual review for privilege in order to be sure that content that could properly be withheld is not inadvertently revealed. Failure to identify a privileged document could jeopardize the interests of the party performing the review, so it is common practice to have highly qualified (and thus expensive) lawyers perform the privilege review. Of course, even experts make mistakes, as the related literature on judging topical relevance clearly indicates. One way of characterizing accuracy is by measuring inter-assessor agreement, which has consistently proven to be lower than one might expect [21, 20]. When searches are done by different users, disagreement might reflect different notions of relevance or, in our application, different ways of reaching decisions regarding privilege. In e-discovery, however, there is a single senior attorney who ultimately certifies the correctness and completeness of the review process, and their interpretation of privilege is thus taken to be authoritative [15].[3]

Carterette and Soboroff have found that when judgments from one person are used to predict system preferences that would be obtained by computing evaluation measures using the judgments of another person, the quality of the prediction can be enhanced by selecting a relatively conservative assessor (i.e., one that has a lower tendency to make a false positive error) as the source of judgments that are the basis for the prediction [5]. This is an intriguing result for our application because in privilege review it is the risk of false negative errors that would generate the greatest concern on the part of the party performing the review.

Recent work has shown that automated techniques that are trained on limited number of human judgments can approach human performance for some review tasks, and in particular on the first-stage relevance review [10]. While there has also been some work on the design and evaluation of automated classifiers to actually perform the privilege review task [7, 9, 19], there is a widely held belief among attorneys that (absent compelling reasons to the contrary such as a need for privilege review at a scale that would otherwise be impractical), reliance on a fully automated classifier for privilege review would incur an undesirable level of as-yet uncharacterized risk. Thus automated classifiers are more often used for consistency checking on the results of a manual privilege review process than as the principal basis for that review. In this paper, we explore a second possible use of the technology. That is, use of automated annotations to (hopefully) improve the accuracy or the cost of a manual review process.

2.2 Document Collection

For our study, we needed a set of documents that we know to be relevant to some request that we might typically see in e-discovery. To train our annotators, we also need a set of

[3]This certification can itself be litigated; in such cases the court would make the authoritative determination.

similar documents that we know to be privileged. We thus need a test collection that contains some relevance and some privilege judgments. One such collection, which we used in this paper, was produced during the TREC Legal Track in 2010.

In the 2010 TREC Legal Track's "Interactive task",[4] one task (Topic 303) was to find *"all documents or communications that describe, discuss, refer to, report on, or relate to activities, plans or efforts (whether past, present or future) aimed, intended or directed at lobbying public or other officials regarding any actual, pending, anticipated, possible or potential legislation, including but not limited to, activities aimed, intended or directed at influencing or affecting any actual, pending, anticipated, possible or potential rule, regulation, standard, policy, law or amendment thereto."* [7] The collection to be searched was version 2 of the EDRM Enron Email Collection, which includes both messages and attachments. The items to be retrieved were "document families," where (following typical practice in e-discovery) a family was defined as an email message together with all of its attachments. Five teams contributed a total of six interactive runs for Topic 303, with each run being a binary assignment of all families as relevant or not relevant. A stratified sample of families was drawn from submitted runs, and 1,090 of those families were judged to be relevant [7]. We have drawn a random sample of 200 of those relevant families for use in our study. Our automated annotation pipeline failed on 12 of those 200 families which lacked a critical field (From, To, or Date), so we removed those 12 families from consideration and randomly split the remaining families into two disjoint sets of 94 families each, which we refer to as D_1 and D_2. We consistently use set D_2 with our Baseline system and set D_1 with our AID system.

In the 2010 TREC Legal Track's Interactive task, a second task (called "Topic 304") was to find *"all documents or communications that are subject to a claim of attorney-client privilege, work-product, or any other applicable privilege or protection"* [7]. Two teams submitted a total of five runs, with each run being a binary assignment of every family as Privileged or Not Privileged. A stratified sample of 6,736 families were marked as privileged or not privileged by experienced reviewers,[5] and prior work has shown that these annotations can be used to train a privilege classifier with reasonable levels of accuracy [19]. A total of seven families from this random sample were, by chance, also present in either D_1 or D_2, and we removed the five that had been judged as Privileged from the set that we used for training our numerical annotators.[6] As Table 1 indicates, this resulted in a total of 932 families annotated as Privileged and 5,799 families annotated as Not Privileged that could be used for training our automated annotators.

Table 1: TREC 2010 privilege judgments for the training-set and the two review sets of families.

	Training	D_1	D_2
Privileged	932	2	3
Not Privileged	5,799	1	1

3. THE AID SYSTEM

Our AID system is a research prototype that is designed to help explore the design space for providing automated assistance to users during privilege review. In this section, we first describe the design of the five types of automated annotators that we have built. We next explain the interface and interaction design of our AID system. A total of 61 of the 94 families in D_1 were reviewed using the AID system by at least one participant during the user study. We characterize the coverage of each of our automated annotators as the fraction of the unique items (people, terms or dates) in those 61 families for which annotations are available.

3.1 Propensity Annotation

Given the central responsibility of people in the definition of attorney-client privilege, a natural choice for an annotator is one that can determine which people have the greatest propensity to engage in privileged communication. As a simplifying assumption, we treat each email address[7] as being associated with a different person. We automatically compute privilege propensity estimates for people using both power iteration and heuristic expansion through egocentric networks [18]. The intuition behind the method is that an initial privilege propensity can be estimated from the annotated training data, and that a fixed point algorithm (similar to PageRank) can then be used to arrive at a somewhat better informed propensity estimate. This aproah implicitly assumes that people who communicate with others who have higher propensity should themselves have somewhat higher propensity). One-hop spreading activation is then used to estimate the privilege propensity for email addresses that were not seen in training. We arbitrarily threshold these (non-negative) propensity values at 0.3 and 0.5 to form three categories that indicate weak, moderate, or strong propensity to engage in privileged communication. Of the 345 unique email addresses that appear as a sender or recipient in at least one of the 61 families viewed using the AID system, 19 were annotated with strong propensity, 34 with moderate propensity, and the remaining 292 with low propensity.

3.2 Person Role Annotation

Propensity annotation is intended to help call a user's attention to a specific person, but actually knowing how to interpret the importance of that person requires additional information. Professional reviewers would typically have information about the roles of specific people (e.g., they might know who the attorneys and the senior executives are), and in complex cases such lists could be quite extensive. The speed, and perhaps the accuracy, of the review process might be enhanced if we could embed that information in the review system. For this purpose, we need a role annotator that can associate each email address with some (generic or specific) version of their job tile. For our

[4]A task in which participants design both a system and an interactive process for using that system

[5]13 of these 6,736 had actually been marked as Unjudged, but during our experiments those 13 were treated as Not Privileged. The effect of this is negligible.

[6]Because of presentation order neither of those Not Privileged documents was seen by any participant in the user study that we describe in this paper.

[7]In Enron's email system, person names were sometimes present in the From, To, Cc or Bcc field in lieu of an email address; we treat such cases as if the person name actually was an email address.

experiments we therefore built a simple role annotator using table lookup. We manually populated this table for 160 of the 1,611 unique email addresses that appear in at least one of the 188 families in either of our two test-sets. We obtained these roles from the MySQL database released by Shetty and Abidi [16], from ground truth produced for evaluating the Author-Recipient-Topic model of McCallum et al [12], from other lists found on the Web,[8] from manual examination of automatically inserted signature blocks in email messages throughout the collection, from public profiles such as LinkedIn, and through manual Web searches. The roles were manually edited for consistency and conciseness. A total of 57 of the 345 unique email addresses that appear as a sender or recipient in at least one of the 61 reviewed families were annotated for person role.

3.3 Organization Type Annotation

When the role of a specific person is sometimes not known, reviewers might benefit in such cases from knowing the type of the organization for which that person works. We therefore used the same lookup table to annotate the organization in such cases. We did this by manually examining the domain name of an email address and then using a current domain name registry, a Web search, or our personal knowledge to label the organization's type, when possible. For example, some messages in the Enron collection are from addresses with the domain 'brobeck.com', and Wikipedia indicates that (at the time) Brobeck, Phleger & Harrison was a law firm. We were able to find what we believe to be reasonable organization types for a total of 21 of the 288 unique email addresses that appear as a sender or recipient in at least one of the 61 reviewed families but that have no person role annotation.

3.4 Indicative Terms

Term unigrams have been reported to be a useful feature set for privilege classification [17], so it is natural to also consider annotating terms. The families in our collection contain many more terms than the email addresses. Hence some approach to feature selection is needed if we are to avoid the display clutter that would result from annotating every term. We perform this feature selection by obtaining the entropy difference for each term. The entropy difference score identifies words that are like words in the Privileged set and also unlike words in the Not Privileged set [13]. To do this, we first tokenize the email message subject field, email message body and extracted text from each attachment for each family in the training set and in the test-set. We then build two unigram language models on these terms (i.e., the unstemmed tokens), one for the 932 families in the training set that were labeled as Privileged, and the other for the 5,799 families in the training set that were labeled as Not Privileged. We then rank each term w present in either of the test-set families using the entropy difference:

$$score(w) = H_p(w) - H_{np}(w)$$

where $H_p(w)$ and $H_{np}(w)$ respectively represent the entropy of the token w in the Privileged and the Not Privileged language models. We then ranked the terms based on the entropy difference score [4]. Negative Entropy difference scores

[8]http://cis.jhu.edu/∼parky/Enron/employees, http://www.desdemonadespair.net/2010/09/bushenron-chronology.html

Figure 1: Indicative terms; Terms with larger font size indicate higher Negative Entropy Difference

indicate terms that are indicative of privilege. We used the top 350 of 3389 (roughly 10%) unique terms with a high negative entropy difference value. Out the the top 350 terms, we annotate 117 terms with the highest negative entropy difference as strongly indicative of privilege, the middle set of 117 terms as moderately indicative of privilege, and the remaining 116 terms as somewhat indicative of privilege.

3.5 Temporal Likelihood

Email communications that focus on the lawsuits often occur during specific time intervals, so it seems reasonable to expect that privileged communication regarding those events might exhibit some predictable temporal variation. We therefore also built an annotator for dates that estimates the likelihood of privileged communication on (or near) that date. To do that, we parse the date field of the email that heads each family in the training set. We then use maximum likelihood estimation with Laplace smoothing to estimate the probability that a family sampled from the set of training families sent on a specific date would be privileged. We calculate that probability estimate as:

$$P(d_i|n_d^x) = \frac{n_{d_i}^p + 1}{n_{d_i}^p + n_{d_i}^{np} + 2}$$

where d_i is the date of the message, $n_{d_i}^p$ and $n_{d_i}^{np}$ are the total number of Privileged and Not-Privileged families sent on d_i respectively. Because TREC performed stratified sampling, designed to oversample potentially privileged families, we expect this to be a substantial overestimate of the actual probability. Nonetheless, we would expect relative values of the estimate to be informative. Of the 55 unique dates on which at least one of the 61 families viewed using the AID system was sent, we are able to annotate temporal likelihood in this way for 36 of those dates.

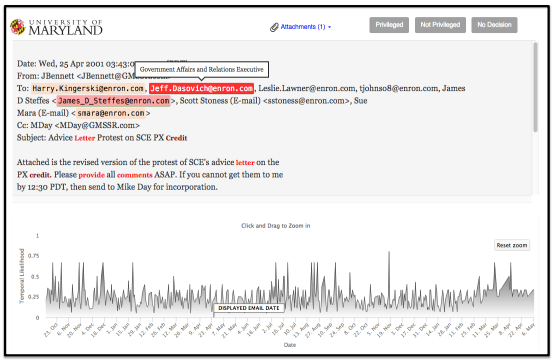

Figure 2: The AID system.

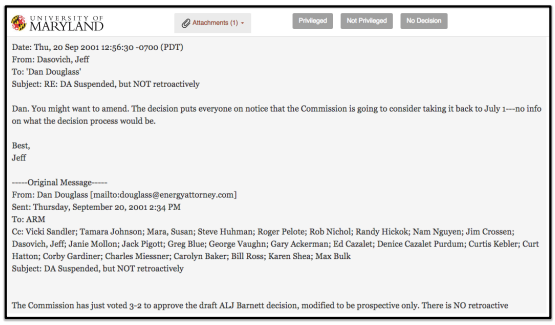

Figure 3: The Baseline system.

3.6 Interface and Interaction Design

Figure 2 shows a screenshot for our AID system. Documents are presented to every user in the same order, and the user must record a judgment (Privileged, Not Privileged, or No Decision) before being shown the next document. They could return to any previously judged document to change their judgment if they chose to do so. Annotations are provided as visual scaffolds during the privilege review process. Whenever a person role or organization type annotation is available, the associated email address is displayed with a red background, and the role or type annotation can be displayed in a manner similar to a "tool tip" (using a graphical control element that is activated when the user hovers the mouse over the shaded area). We shade the background with variations of the color red to indicate the propensity category (darker red for strong propensity, lighter red for moderate propensity, very light red for all other cases in which role or type information is available).[9] On average (across the 61 viewed families), 58% of the email addresses appearing as senders or recipients had a role or a type annotation available (55% for person role, 3% for organization type). About two-thirds of those cases in which role or type annotation were available, were displayed with shading indicating strong or moderate propensity.

The display of terms that are indicative of privilege in the subject line, email message body, or attachments follows a similar pattern, but by altering the color of the typeface rather than the background. For example, the term "credit"

[9]Low propensity addresses for which no role or organization type information is available have no background shading.

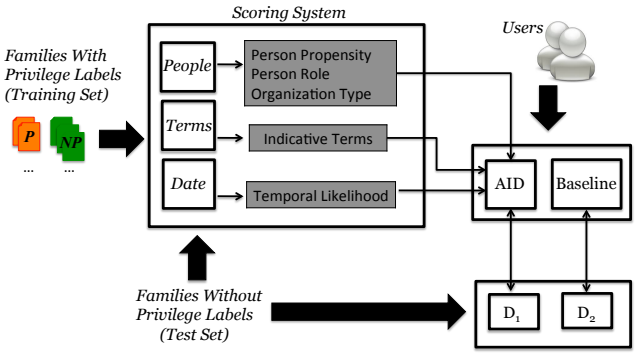

Figure 4: User study overview

is rendered in the darkest shade of red[10] in Figure 2, thus indicating it was strongly indicative of privilege. On average (across the 61 viewed families), 2% of all term occurrences are highlighted.

Temporal likelihood is plotted as a connected line plot, with date as the horizontal axis and temporal likelihood as the vertical axis. This has the effect of visually performing linear interpolation of temporal likelihood for dates on which that likelihood can not be computed directly. The displayed date range can be reduced (by a click and drag zoom-in functionality) by the user for finer-grained display.

Figure 3 shows the user interface of our Baseline system. As can be seen, the only differences from the AID system are that none of the annotations are present, and the omission of the temporal likelihood plot permits more of the content to be displayed. Both the systems log the time, family ID, user ID and judgment (Privileged, Not Privileged, or No Decision) for each reviewed family.

4. USER STUDY DESIGN

The principal goal of our user study was to determine whether any of our system-generated annotators could help the users to perform the review task more quickly, more accurately, or both. A secondary goal was to determine whether there were usability issues with our current interface design that might adversely affect our ability to determine the effects of specific annotators. A third goal was to use our current AID system design as an artifact around which we could discuss specific as-yet unimplemented capabilities that experts might believe would provide useful support for the task. In this section we describe the design of the study, including a brief description of a pilot study that we performed to finalize our procedures, a description of the participants in our study, and a description of the procedures that we ultimately used with each of those participants.

4.1 Pilot Study

We chose to save our limited pool of qualified participants for our actual study. Thus, for the pilot study in which our only goal was to wring out our system design and study procedures, we recruited two participants, neither of whom had a law degree nor previous privilege review experience.

[10]We chose to use the same color gradations for terms and email addresses to simplify training, but the question of optimal color choices merits further investigation.

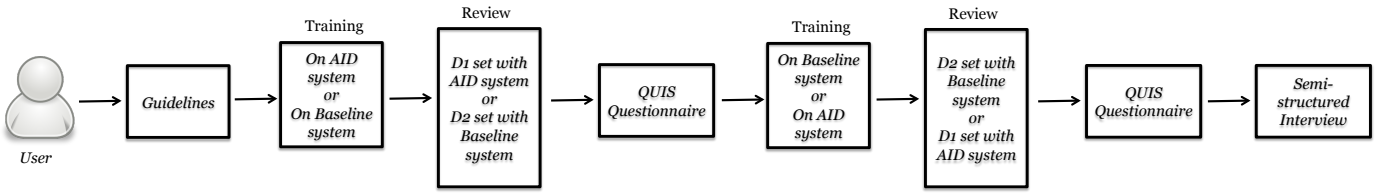

Figure 5: User study procedure.

Each pilot study participant completed the full study as depicted in Figure 5, including completing questionnaires and participating in a semi-structured interview. As in our actual study, we counterbalanced the presentation order, with one pilot study participant using the AID system first and then the Baseline system; the other participant used the two systems in the other order. We made three consequential changes as a result of our pilot study: (1) The organization type annotator was built, based on a suggestion from one pilot study participant, (2) some aspects of the brief written summary of the study procedure that we presented to participants during training were clarified, and (3) the semi-structured interview questions were improved.

4.2 Participants

In recruiting participants for our study, we initially sought people who were (in preference order) (1) a practicing attorney, (2) a law school graduate, (3) a law school student (preferably in their 3rd year), or (4) a law librarian. As it happened, we were able to recruit a total of six participants from the first two groups, which we judged to be adequate for the comparisons we wished to make, so we limited our study to those six participants.[11]

Two of the six were senior attorneys employed by law firms with a current e-discovery practice. These senior attorneys are experienced litigators who have extensive experience conducting both relevance and privilege review for email using commercial Technology Assisted Review (TAR) tools.[12] We refer to these senior attorneys as S_1 and S_2.

The remaining four participants were law school graduates. Two of the four had prior experience conducting relevance and privilege review using commercial TAR tools, but neither was currently working in an e-discovery practice; one of the two is a graduate student in another discipline, the other is an intellectual property attorney. We refer to this pair of experienced reviewers as E_1 and E_2. By coincidence, E_2 had experience working as a reviewer during the original Enron litigation.

The remaining two participants had experience conducting e-discovery reviews some time ago, principally on paper, but neither had experience using current TAR tools. One was a retired attorney, the other was currently a faculty member in another discipline. We refer to these (TAR) inexperienced reviewers as I_1 and I_2. I_2 had little direct experience using computers.

[11]One of our goals was to identify usability issues, and Nielsen claims that most usability problems can be identified with a maximum of five users [14]. More users would have resulted in greater statistical power for our quantitative comparisons, but at this point our system design is not yet sufficiently mature to merit looking for small effect sizes.

[12]Tools like Recommind, Nuix, kCura, etc.

Table 2: Task Order

Participant	Task Order
S_1	Baseline — AID
S_2	AID — Baseline
E_2	Baseline — AID
E_1	AID — Baseline
I_2	Baseline — AID
I_1	AID — Baseline

4.3 Procedure

Participants were given the opportunity to choose the time and the location of their session; 4 came to our lab, and for the other 2 we went to their office. As both the AID and Baseline systems are Web applications, participants were free to use their own computer if they wished; S_1 and S_2 did so.

Because we chose a within-subjects design, we needed to present different families in the two conditions in order to avoid memory effects. We did not have a sufficient number of participants to counterbalance document-system interactions or presentation order effects, so we elected to present the same documents, in the same order, to every participant who used the same system. The presentation order for each of the two collections (D_1 and D_2) was thus randomized once and then frozen. Thus what we refer to as a "system" in our study (i.e., AID or Baseline) is actually an invariant combination of the system, the collection assigned to that system, and the presentation order assigned to that collection.

The participants were randomly assigned to one of the two conditions, either AID first or Baseline first, as shown in Table 2.[13] This counterbalancing was intended to control (to some degree) for learning and fatigue effects. Thus we have two independent variables; user type (S, E, or I) and system.

Figure 5 summarizes the study procedure for one of the six single-participant sessions. Each participant completed the study in about two hours, with a 10 minute break at the end of the first hour. Participants were given an overview of the review task and were asked to read a written description of the study that we provided before signing a consent form. Each participant then received a 5 minute tutorial on the first system they would use, presented by the investigator (the first author of this paper), in which the different parts of the system were demonstrated using a few example fam-

[13]The E and I participants were assigned randomly without regard to the pairings that we ultimately made because prior to the study we did not have enough insight into the background of our participants to recognize the optimal pairings. We are thus fortunate that our post-hoc assignment of participants to comparable pairs turned out to be counterbalanced.

Table 3: Contingency table for annotations of the same families by S_1 and S_2)

	S_1: Privileged	S_1Not Privileged	S_1: No Decision	S_1: Not Seen
S_2: Privileged	15	7	0	5
S_2: Not Privileged	5	62	0	16
S_2: No Decision	6	12	1	3
S_2: Not Seen	0	1	0	75

†There was one family that was skipped in sequence by chance by S_2; but was not skipped by S_1.

ilies drawn from outside either test collection. Participants were then offered an opportunity to practice using those example families until they felt ready. The participant then logged in to the review system (with an anonymized user-ID that we provided) and perform the review task for 30 minutes without interruption. The review task required each participant to sequentially read and mark each family in the test set associated with their system (D_1 for AID, D_2 Baseline) as either Privileged, Not Privileged or No Decision. Participants were able to refer to their copy of the written task description as they performed the task. At the end of 30 minutes (which always occurred before the users had recorded decisions for every family in a test set), participants were presented with a short usability questionnaire for that system that we created by tailoring the Questionnaire for User Interaction Satisfaction (QUIS) [6] to use 9-level scales for each question. After a 10-minute break, they were then trained on the second system (using the same procedure as before) and asked to log in to that system using their anonymized user-ID and complete a second 30-minute review of the families in the test collection associated with that system. After a second QUIS questionnaire for that system, participants were asked a set of semi-structured interview questions, some of which called for subjective judgments. The goal of the semi-structured interview was to obtain the participant's perspectives on the degree to which they found specific types of annotations to be useful, and to characterize the extent to which they found aspects of the task design to be mentally demanding or frustrating. With the participant's permission, this brief (5 minute) semi-structured interview was recorded to facilitate later analysis. At the conclusion of the session, each participant received a US$25 in cash.

5. RESULTS

In this section we first focus on quantitative results for accuracy and speed. Following that we contextualize these results from qualitative results from our interview and from our usability questionnaire. We then draw insights from each of these analyses to discuss what we see as the most important conclusions that can be drawn from this study.

5.1 Selecting a Benchmark for Evaluation

If we are to make any useful statements about the accuracy of a privilege review, we must first select an informative set of judgments as benchmark against which accuracy can be measured. This benchmark judgments need not be perfect for the resulting measures to be informative, but we will have the greatest confidence in our results if we select the best available benchmark judgments. Thus it is natural to begin by characterizing the results from the two senior attorneys, since we would expect their judgments to be natural candidates as a benchmark.

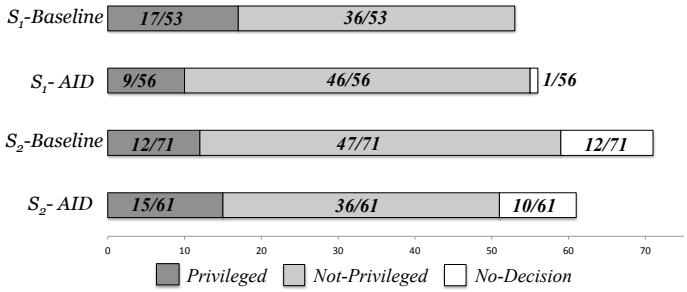

Figure 6: S_1 and S_2 Judgments by type

Figure 6 shows the number of judgments of each type made by S_1 and S_2 for each of the two conditions. As can be seen, S_2 is somewhat faster than S_1 (making 33% more judgments in the same 30 minutes in the Baseline condition, and 9% more in the AID condition). S_2 records many more No Decision judgments (22 for S_2 vs. 1 for S_1).[14] As Table 3 shows, senior attorney S_1 marked a total 15+5+6=26 families as Privileged while S_2 marked a total of 15+7+5=27 families as Privileged. Among the families seen by both senior attorneys (using either system), 15 families were marked as Privileged by both. Computing chance corrected inter-annotator agreement between S_1 and S_2 using Cohen's Kappa (κ) yields 0.68, a value that Landis and Koch [11] characterize as "substantial." Indeed, given the class prevalence in our test sets, chance agreement would be 0.57, making very high levels of κ difficult to achieve [3].

As Table 1 showed, TREC 2010 Interactive Task Topic 304 privilege judgments are available for seven of the families in our test set. Of those seven, 5 were Privileged and 2 were Not Privileged. Of the 5, three families were adjudicated by the Topic Authority (a senior attorney whose judgments were authoritative) who was responsible for providing guidance and adjudicating disputes. Out of the three Privileged families adjudicated by the TREC Topic Authority, two were reviewed by both S_1 and S_2. S_1 agreed with the Topic Authority on one of the two families by marking one of the two families as Privileged while the other as Not-Privileged. S_2 never agreed with the Topic Authority. S_2 marked one of the two families as Not Privileged (the same family marked as Not Privileged by S_1) and the other was marked as No Decision. Comparisons on two judgments is not sufficient to determine whether the two senior attorneys in our user study are (1) generally more inclined to judge documents as Not Privileged than the TREC Topic Author-

[14]Participants could mark a family as No Decision when a clear distinction between Privileged and Not Privileged could not be made on the email message or any of its attachments.

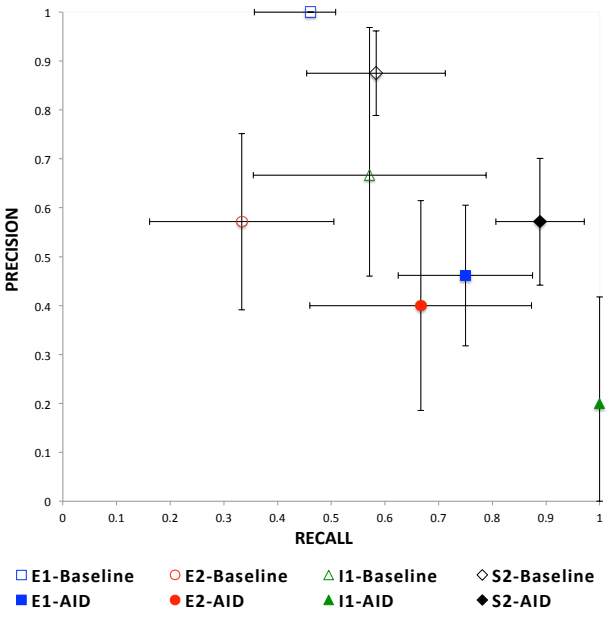

Figure 7: Evaluation with S_1 judgments as Benchmark.

ity would have been (2) generally inclined to agree with each other, but we can say that there is no evidence to refute such a claim.

From this analysis, either senior attorney could reasonably be chosen as a benchmark against which the other participant's judgments could be measured for accuracy. However, because S_2 left 19 families unjudged and skipped reviewing 1 family throughout the review sequence and all 24 of the families that were not seen by S_1 were late in the review sequence, a larger number of useful judgments are available from S_1. We therefore use judgments from S_1 as a benchmark for evaluation. We evaluate participants on the basis of precision and recall estimates that we report in Figure 7.

5.2 Accuracy

Figure 7 shows the privilege review effectiveness of S_1, E_1, E_2, and I_1 for the Baseline and AID conditions, evaluated as if the judgments by S_1 were the ground truth. We calculate point estimates for precision and recall using only the cases judged as Privileged or Not Privileged by both S_1 and by the participant whose decisions are being evaluated (i.e., we omit No Decision and Not Seen cases from both). Because we are comparing estimates for different sets of documents, we also show the 95% confidence intervals for recall and for precision, computed using the standard approximation method described by Agresti et al. [2]. Results for I_2 are not shown because after removal of the 21 No Decision judgments recorded by I_2 there were only 7 families judged by I_2 (3 in the AID condition, 4 in the Baseline condition), a number insufficient for useful estimation of intervals.[15]

From Figure 7 we can conclude that there is a consistent and statistically significant improvement in recall when the review task is performed using our AID system for all four participants (S_2, E_1, E_2, I_1).[16] This improvement is, however, accompanied by a statistically significant reduction in precision for three of the four participants. Using S_2 as a reference to evaluate S_1, E_1, E_2, and I_1 (not shown) yields results, with statistically significant improvements in recall in 1 of 4 cases and statistically significant decreases in precision in 2 of 4 cases. Since the principal goal of our AID system is to avoid inadvertent disclosures, this consistent bias in favor of recall (i.e., in avoiding false negatives), regardless of which senior attorney we select as a reference, is well in line with that goal.

5.3 Speed

To characterize the effect of the choice of system on review speed, we computed the number of families reviewed by each participant in 30 minutes using the Baseline and the AID systems, observing little difference in the means (averaging 40.1 families for the AID condition and 43.6 families for the Baseline condition).[17] A paired t-test found no detectable difference in average review speed across the two conditions ($p > 0.38$). From these results we conclude that there is no indication that our AID system results in faster review, and indeed it is possible that our AID system might result in marginally slower review.

5.4 Usability

Table 4 summarizes participant responses to six of the seven QUIS questions (a seventh question, about layout, evoked no useful differences in the responses). Five of the six participants assigned a higher rating to the overall experience with the AID system than with the Baseline system (and the sixth participant noted no difference). All six participants gave more positive scores to the AID system than to the Baseline system in response to the question about adequacy of the displayed information. Person highlighting was reported to be useful (to at least some degree) by five of the six participants, whereas term highlighting and the date graph were each reported to be useful to some degree by only two of the six participants.

5.5 Usefulness

During the semi-structured interview session, we asked each participant which type of system-generated annotation they found to be most useful; five of the six named person annotation. The following excerpts are representative of responses that participants gave to our open-ended questions.

"I think having the role or type information in-line on the user interface was very helpful. All I had to do was to hover over the name instead of looking it up on a sheet of paper as we normally do." — S_1

"I would honestly like the people highlighting concept much more if it would give me more information about the metadata. Having information about the domain addresses of people who are not Enron employees is one such information." — S_2

"The presence of highlighted people made me look into the documents more carefully in non obvious cases for the presence of potentially privilege content. It help me to make a

[15] All 7 were judged as Privileged, suggesting that participant I_2 may have intended to record judgments of Not Privileged and instead incorrectly selected No Decision. It was participant I_2 who had only limited personal experience using computers.

[16] We consider a difference to be statistically significant if each point estimate lies outside the 95% confidence interval for the other condition.

[17] Data from participant I_2 is omitted from this analysis.

Table 4: QUIS Summary(SA=Strongly Agree, A=Agree, MA=Moderately Agree, SD=Strongly Disagree, D=Disagree, MD=Moderately Disagree; NF=Neutral Feedback, blank indicates not applicable; BL=Baseline).

	S_1 BL	S_1 AID	S_2 BL	S_2 AID	E_1 BL	E_1 AID	E_2 BL	E_2 AID	I_1 BL	I_1 AID	I_2 BL	I_2 AID
Overall review experience	NF	Good	Bad	Good	Good	Great	Good	Good	Bad	Good	Fair	Good
Information provided was adequate	D	A	SD	MD	A	SA	A	SA	SD	MD	SD	A
People highlighting was useful		SA		A		A		SA		SA		NF
Term highlighting was useful		SA		NF		MA		NF		NF		SD
Date graph was useful		NF		D		D		MA		D		A
Use of colors was logical		A		A		SA		NF		A		A

filtering decision about which document need more attention. The highlighting helped me to be quicker." — E_1

"I think the trickiest part was to review the document when the information about a sub-set of the people was missing. For example, if there were 6 people and we have information about 3 of them but not the other 3, it is hard to predict who the other players are." — E_2

"I think the highlighting of the people was useful to do the review; the highlighting of the terms were less useful because almost all emails contain the same boilerplate language and the term highlights did not provide much information; and about the dates, I did not feel the need to use the date information displayed on the graph." — I_1

"The ideas presented in the AID system are good, however the information provided was sometimes confusing to me. The role and type information provided was useful but the term highlighting was distracting; mainly because the highlighted terms did not make sense to determine privilege and I lost my faith on the terms." — I_2

5.6 Discussion

Our quantitative results clearly indicate that our AID system resulted in a greater ability to detect privileged documents, and that that improvement is more than would be expected from chance variations resulting from the presence of different families in the test collections for each condition. The QUIS and our semi-structured interviews provide consistent and well triangulated support to our belief that our annotation of people (or, more specifically, of email addresses) is principally responsible for this improvement. Of the three ways we annotate people (for propensity, for person role, or for organization type) we have the strongest evidence for a claim that role and organization type annotation was believed by our participants to be useful; we do not have sufficient evidence to separately identify the effect of propensity annotation. Neither our present implementations of term highlighting nor the date graph were often commented on favorably by the participants. From these observations, we conclude that our current AID system achieves its principal objective of helping to avoid inadvertant disclosure, that further study is needed to separately analyze the value of propensity annotation, that the value of term annotation has not yet been shown, and that further refinement of our approach to date annotation will not be among our highest near-term priorities. We base this last conclusion in part on the following comment by S_1, who said *"Date information could be helpful during responsiveness review. But for privilege review, it is less likely to be useful"*.

We were somewhat surprised by the magnitude and consistency of the drop in precision that accompanied the increases in recall that we observed from the use of our AID system. In privilege review, low precision could result in incorrectly withholding some families that should properly have been turned over to the requesting party. Perhaps such cases might be discovered and corrected in a second stage of privilege review, but a two-stage review process would naturally lead to higher costs. Future work aimed at understanding the reason for the reduction in precision will thus be a high priority. Moreover, trade-offs between recall and precision are natural, so it may be that similar results might be obtained in other ways (e.g., by providing financial incentives based on the number of privileged documents found). In future work it will therefore be important to develop task-tuned utility measures that account for the relative importance of recall and precision for the privilege review task and to develop study designs in which recall at comparable levels of precision can be studied.

Our participants made some suggestions for improvements that might be made to our AID system. One useful suggestion was to consider highlighting multi-word expressions that are indicative of privilege, rather than only single-terms as our present system does. Another useful suggestion was to consider augmenting our role annotations with an opportunity to drill down to learn more (e.g., date assigned to that role, previous roles, or supervisory relationships). In future work we are interested in exploring the potential for viewing privilege review as a structured collaboration task, and when we asked about this several of our participants (three of the five who we asked) indicated that system support for collaboration might be of interest for privilege review.

6. CONCLUSIONS & FUTURE WORK

In recent years, system support for relevance review has been the center of the action. As the technology-assisted review tools are deployed and adopted, it is natural to expect larger cases to be tackled, with a concomitant increase in the number of relevant documents that require privilege review being one predictable consequence. It therefore seems timely to begin to think seriously about how the tools that are ultimately built to support privilege review will differ from those that support relevance review. Our design thinking for that problem began with the idea that modeling attributes of people such as their roles and their propensity to engage in privileged communication might be particularly important for the privilege review task, and our results provide support for that belief. Our results also indicate that dates, while unquestionably important for relevance review, may be of less value for privilege review (at least in the way we are doing things now). We have noted that the increases in recall that we observed were often accompanied by substantial declines in precision, and further study will be needed to better characterize this effect and to control for it in future

experiments. Contrary to our expectations, we also noted no evidence of improvements in review speed, although of course even our most expert participants were novice users of the particular interface that we presented them with. In future work we may therefore consider longitudinal studies that would allow us to see how the same users behave at different points in their personal learning curve.

One obvious next step will be to run small-scale studies to tune specific components (e.g., what types of multi-word expressions should be considered for highlighting? how many terms or multi-word expressions should be highlighted? how many categories of term highlighting are useful?). Studies along those lines might ultimately lead to test collections that could be used as a basis for tuning and evaluating specific system components; for that we will also need to give thought to intrinsic measures for evaluating the performance of individual components. Another productive research direction would be to explore whether we might productively use surrogates for attorneys in some early studies. Would law students be suitable? Law librarians? Crowd-sourcing services such as Mechanical Turk? Surely we can go some distance in this direction; the key question is how far can we productively go. Another important direction for future research will be to unpack the human decision process a bit to study how it is that attorneys actually make privilege judgments in specific cases; that could be a productive source for design inspiration. Further study is also needed on the question of the degree to which use of such a system over time could (or should!) engender trust on the part of reviewers who themselves bring expert knowledge to the task. Yet another direction we might go is to consider what happens when things change. Manually constructing lists of person roles and organization types was a useful expedient for this study, but actual cases evolve dynamically, and they do so within organizations that are themselves often dynamically restructuring. There is, therefore, great scope for integrating techniques that can learn to make useful inferences from raw content.

Finally, we should note that this work could be extended in other settings where search amidst sensitive content is inevitable. On December 8, 2009, President Barack Obama of the United States wrote *"Information maintained by the Federal Government is a national asset. My Administration will take appropriate action, consistent with law and policy, to disclose information rapidly in forms that the public can readily find and use. Executive departments and agencies should harness new technologies to put information about their operations and decisions online and readily available to the public."* Perhaps paradoxically, the caveat "consistent with law and policy" means that before his nation can reap the full benefits of the information that can and should be disclosed, his government will need the ability to affordably separate that vast trove of immediately useful information from that which – at least for now – can not and must not be disclosed. That problem bears an uncanny resemblance to the challenge of privilege review in e-discovery, and indeed progress on one could well lead to progress on the other.

7. ACKNOWLEDGMENTS

This work has been supported in part by NSF award 1065250. Opinions, findings, conclusions and recommendations are those of the authors only and may not reflect NSF views.

8. REFERENCES

[1] Armstrong v. Bush, 1989.

[2] A. Agresti and B. A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 1998.

[3] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 2008.

[4] A. Axelrod, X. He, and J. Gao. Domain adaptation via pseudo in-domain data selection. In *EMNLP*, 2011.

[5] B. Carterette and I. Soboroff. The effect of assessor error on IR system evaluation. In *SIGIR*, 2010.

[6] J. P. Chin, V. A. Diehl, and K. L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *SIGCHI*, 1988.

[7] G. V. Cormack, M. R. Grossman, B. Hedin, and D. W. Oard. Overview of the TREC 2010 legal track. In *TREC*, 2010.

[8] E. S. Epstein. The Attorney-Client Privilege and the Work-product Doctrine. American Bar Association, 2001.

[9] M. Gabriel, C. Paskach, and D. Sharpe. The challenge and promise of predictive coding for privilege. In *ICAIL, DESI V Workshop*, 2013.

[10] M. R. Grossman and G. V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Rich. JL & Tech.*, 2011.

[11] J. R. Landis and G. G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374, 1977.

[12] A. McCallum, A. Corrada-Emmanuel, and X. Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. 2005.

[13] R. C. Moore and W. Lewis. Intelligent selection of language model training data. In *ACL*, 2010.

[14] J. Nielsen. *Usability engineering*. Elsevier, 1994.

[15] D. W. Oard and W. Webber. Information retrieval for e-discovery. *Foundations and Trends in Information Retrieval*, 2013.

[16] J. Shetty and J. Adibi. The Enron email Dataset; Database schema and brief statistical report. *ISI technical report, USC*, 2004.

[17] J. K. Vinjumur. Evaluating expertise and sample bias effects for privilege classification in e-discovery. In *ICAIL*. ACM, 2015.

[18] J. K. Vinjumur and D. W. Oard. Finding the privileged few: Supporting privilege review for e-discovery. *ASIS&T*, 2015.

[19] J. K. Vinjumur, D. W. Oard, and J. H. Paik. Assessing the reliability and reusability of an e-discovery privilege test collection. In *SIGIR*. ACM, 2014.

[20] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 2000.

[21] W. Webber. Re-examining the effectiveness of manual review. In *SIGIR Workshop*, 2011.