Finding the Privileged Few: Supporting Privilege Review for E-Discovery

Jyothi K. Vinjumur
College of Information Studies
University of Maryland, College Park-20742
jyothikv@umd.edu

Douglas W. Oard
College of Information Studies & UMIACS
University of Maryland, College Park-20742
oard@umd.edu

ABSTRACT
As typically conceived, the goal of information retrieval is to help people find what they seek. However, when sensitive content is intermixed with what is sought, it can be equally important to help protect what should not be seen. This paper focuses on one such case, the protection of content that is subject to a claim of attorney-client privilege when sharing evidence incident to civil litigation, a process called e-discovery. Although lawyers have been quick to embrace text classification techniques for finding relevant evidence, the stakes involved in inadvertent disclosure of privileged content create reluctance to trust any fully automated technique to accurately recognize content that can properly be withheld. Rather than starting with privilege classification as a goal, this paper proposes a modest first step for building tools that can help lawyers to make faster and more accurate privilege judgments by scoring the importance of specific email addresses to determine their propensity to engage in privileged communication. Both recursive and heuristic techniques are used to estimate that propensity score, ultimately resulting a coverage of 94% of the email addresses.

Keywords
E-Discovery, Privilege Review, Scoring Algorithm.

MOTIVATION & BACKGROUND

Civil litigation is a legal dispute between two or more parties in which money, but not life of liberty, is at stake. In civil litigation in the United States, parties to a lawsuit have the right to request relevant evidence from each other, a process known as "discovery." (Discovery in this context refers only to the process in civil litigation.)

Parties can, however, assert a privilege that allows them to withhold relevant evidence in specific circumstances, such as when a lawyer has provided legal advice to a client. Such privileges exist to foster specific desirable outcomes, such as the fully informed advocacy that open communication between lawyers and clients can support. Discovery (particularly in the context of electronically stored digital content, so-called "e-discovery") thus poses a hydra-headed challenge to information retrieval systems: we must initially find the evidence that has been requested, and among that requested evidence we must identify (and withhold) that which is privileged. Failure to find the requested evidence jeopardizes the interests of the requesting party; failure to identify privileged evidence jeopardizes the interests of the responding party.

Unsurprisingly, neither side is particularly eager to trust automated systems to make these decisions on their behalf. Requesting parties have been the first to agree to automation, in part because the volume of potential evidence that must be searched simply overwhelms even well resourced manual processes; searching billion-document collections is becoming increasingly common. As a consequence, text classification techniques have gained prominence in e-discovery. The set of documents that have been judged responsive to a request is typically far smaller, however, so responding parties still prefer to examine each document for privilege before it is turned over to the requesting party. Automated classification techniques can still be used in such cases (e.g., as a final check to find decisions that appear to the classifier to be inconsistent), but as other costs have come down, the manual document-by-document privilege review process has been one of the largest cost drivers in complex cases.

Although many types of documents can be important in e-discovery, email is prominent because much of the activity of an organization is ultimately reflected in the emails sent and/or received by its employees. For this reason, email provides an excellent environment to initially develop techniques to improve the productivity and accuracy of privilege review. In this paper, we explore some aspects of the design of such a system using Enron emails that were annotated for privilege during the TREC 2010 Legal Track.

Privilege in E-discovery

Privilege in general is a right given to an individual that provides protection against the involuntary disclosure of information. Attorney-client privilege in particular aims to protect the information exchange between "privileged persons" for the purpose of obtaining legal advice.

ASIST 2015,November 6-10, 2015, St. Louis, MO, USA.

Privilege does not arise simply because privileged persons communicate; it can only be claimed when the content of the communication merits the claim. We therefore need two types of models: one for "privileged persons" and one for recognizing "privileged content." In this paper, we focus on the question of identifying people (or their email accounts) with high propensity to engage in privileged communication, leaving content analysis to future work.

Privileged persons include (Epstein, 2001):

- the client (an individual or an organization),
- the client's attorney,
- communicating representatives of either the client or the attorney, and
- other representatives who may assist the attorney in providing legal advice to the client.

Moreover, including people in an otherwise privileged communication who have no business relation to the matter being discussed can waive the privilege. We are, therefore, interested in not just who exhibits a high propensity to engaged in privileged communication, but also in who exhibits a low propensity.

Test Collection

During the 2010 TREC Legal Track (Cormack, 2010), one of the tasks (Topic 304) asked participants 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." (Although the annotations created by the 2010 TREC Legal Track reflect Attorney-Client privilege, the attorney work product doctrine, and other protections as well, we motivate our design by the nature of Attorney-Client Privilege.) The collection to be searched was version 2 of the EDRM Enron email collection, which includes both messages and attachments. The items to be classified were "document families" that consisted of an email message together with all of its attachments.

In the 2010 TREC Legal Track (Cormack, 2010), the first-tier assessors were employees of a legal services firm that routinely provides document review services. An assessor's judgment on any family could be escalated for adjudication by a single senior attorney, referred to as the Topic Authority (TA), under three conditions: (1) a team could appeal the decision of an assessor in cases where they believed a clear error had been made – a total of 237 appeals out of 6,766 total annotated families were received; (2) 730 families were sampled for dual assessment by a second first-tier assessor — off the 730, 76 families that received conflicting judgments were escalated for adjudication; and (3) a stratified sample of 223 non-appealed and non-adjudicated families was drawn in order to characterize the error rate of first-tier assessors (by treating the TA's judgments as a gold standard). We thus have a total of 536 families that were annotated by the TA, creating the Adjudicated Set (AS).

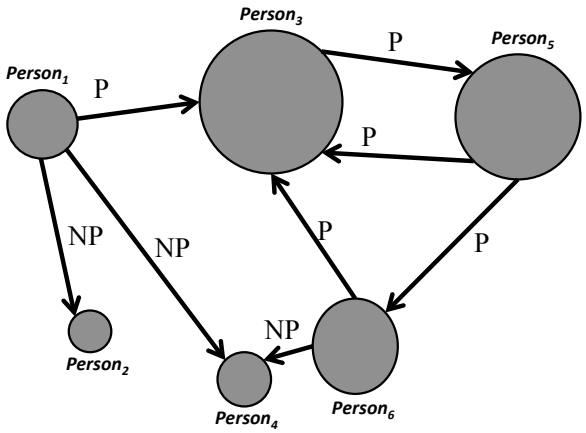

Figure 1. An Email Network Example

The remaining 6,230 families that were annotated only by first-tier annotators constitute the Non-Adjudicated Set (NAS). Our earlier work (Vinjumur, 2015) compared the effects on classifier training using AS (expert) and NAS (non-expert) families and showed that sampling bias could affect the classifiers performance. In this paper, we ask a similar but more specific question; we ask whether the two training sets (AS or NAS) can help to derive useful privilege propensities.

APPROACH

The key intuition behind our approach is that, an email message sent or received by a person (e.g. $Person_3$) has a higher probability of being involved in privileged communication if that person frequently communicates with other people ($Person_5$, $Person_6$, etc.) who themselves have a higher probability of being involved in a privileged communication. Figure 1 illustrates this idea. As shown in the figure, the node $Person_3$ in the example email network has multiple privileged (P) email exchanges with the node $Person_5$ which in-turn has privileged email exchange with $Person_6$. The privilege propensity of node $Person_3$ depends not only on the emails sent/received by $Person_3$, but also on the email traffic of all the nodes $Person_3$ communicates with. Thus we define "propensity" as a measure of the degree to which we expect a person to engage in privileged communication. It is a number between 0 (low propensity) and 1 (high propensity).

We use power iteration method to calculate the propensity scores over a large set of users based on a limited amount of training data. We utilize the AS and NAS annotations separately to compute the propensities and compare the results. Thus, given the labels of the edges, the task is to assign a score to the nodes that depends on the edge labels. We start by computing a privileged communication probability value that is associated with each edge in the graph as a prior, using the network information from labeled families in the training set. We then use the idea behind the weighted PageRank technique to score the propensity for each person (Xing, 2004).

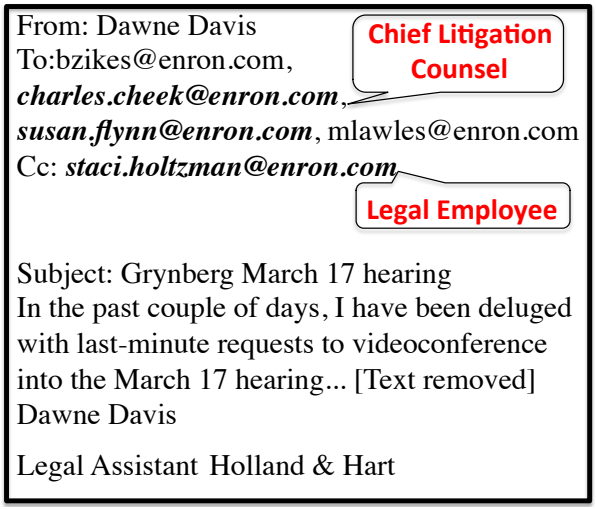

Figure 2. Privileged Email Example

We define $w[Edge(x,y)]$ as the edge weight between x and y given the label of each communication edge as:

$$w[Edge(x,y)] = \sum_{e \,\varepsilon\, E_{train}} \frac{n(x,y)_{e_p}}{n(x,y)_{e_p} + n(x,y)_{e_{np}}}$$

where E_{train} is a set of labeled edges used in training set; $n(x,y)_{e_p}$ and $n(x,y)_{e_{np}}$ are the number of edges labeled as privileged and not-privileged respectively, with x as sender and y as the recipient. The weight of $Edge(x,y)$ indicates the privilege probability between the two people. To score the individual nodes, we use these weighted edges in the graph as an input to a power iteration algorithm to obtain the "propensity score" or PR_{score} for each node/person using:

$$PR_{score}(x) = (1-d) + d \sum_{v \,\epsilon\, E_x} \frac{PR_{score}(v)}{N_v}$$

where d=0.85[1] is the dampening factor; by fixing the value of d to 0.85, we assign 15% privilege likelihood for persons with no prior labeled privileged communications, E_x is the set of edges where x is the recipient; and N_v is the total number of edges where v is the sender.

Scoring Algorithm

Given the PR_{score} of each person seen in the labeled training set families, the final step of our person-scoring algorithm is to calculate the PR_{score} of each person seen in the test-set. While training on families from the AS set, we observe only 30% of the senders or recipients of test-set emails associated with a PR_{score} (32% in case of training on families NAS). To estimate the propensity for people who are not present in the training set, we leverage each unknown persons' egocentric communication network, ultimately increasing the number of people to whom we can assign a propensity score to 93% of senders and recipients in the test set when training on AS families and 94% when training on NAS families. Figure 2 shows an example family where none of the 6 persons are seen in the training set. However, our missing person algorithm scores 3 of the 6 (shown in bold font).

To calculate the propensity score for each person in the test set, our algorithm follows two steps:

1. Common Person Scoring: We obtain a set of common persons (persons seen in both the training and test sets) $Common_a$. For each person i in the test set, if $[i \,\epsilon\, Common_a]$, then we use the $PR_{score}(i)$.

2. Missing Person Scoring: For each person i in the test set, if $[i \notin Common_a]$, we take the approach described in the Algorithm 1. For each person in the test graph who is not seen in the training graph, we exploit the person's network information. If the missing sender is connected to one or more recipients who are seen in the training graph, we assign the average of recipient's node scores as the missing sender score. However, if the sending person is connected to only missing recipients, we assign the sender the average of all PR_{score} values in the training graph. We take this conservative approach to scoring missing persons, because we do not want to mislead the user by providing a zero propensity score when we are actually simply unsure about the propensity.

Algorithm 1 Missing Person Score Algorithm

Input:
$Graph_{test}$
$PRdictionary_{score}(A_{train})$
$uniqueNodes_{test}$

procedure $GetMissingNodeScores_{test}$
  $rankDictionary \leftarrow sort(PRdictionary_{score}(A_{train}))$
  $uniqueNodeScoreDict \leftarrow NULL$
  for <each edge(s,r) in $Graph_{test}$ > **do**
    if s in $uniqueNodes_{test}$ **then**
      $sum \leftarrow zero$
      if **then** r in rankDictionary.keys()
        $sum \leftarrow sum + d[r]$
      else
        $sum \leftarrow sum + [(min(d.values()) + max(d.values())) \div length(d)]$
      end if
    end if
    $score_n \leftarrow sum \div num(recipients)$
    $unqNodeScoreDict[n] \leftarrow score_n$
  end for
  return $unqNodeScoreDict[n]$
end procedure

RESULTS

We used the 536 TA annotated families as "AS" training set. Since 45% of those 536 families were judged as privileged, in the "NAS" training set we randomly selected an equal number of families from NAS with roughly the

[1] In this paper we fix the damping factor "d" to the empirically chosen value used in PageRank (Boldi, 2005).

Propensity	Person Name	Role
0.80	ariel.cerantola	No Information
0.64	mark.taylor	Counsel
0.64	gary.taylor	Director-Risk Management
0.62	jeff.blumenthal	In-house Tax Lawyer
0.62	marcelo_cosma @ml.com	Counsel at Merrill Lynch
0.61	nidia.mendoza	Credit Department
0.61	sara.shackleton	Lawyer
0.49	brant.reves	Enron Freight Markets
0.46	jason.peters	ENA Counsel
0.46	trevor.mihalik	Chief Financial Officer

Table 1. Top 10 Propensities, Trained on AS

Propensity	Person Name	Role
0.87	mark.taylor	Counsel
0.87	gary.taylor	Director-Risk Mgmt
0.81	richard.shapiro	VP Regulatory Affairs
0.48	dan.j.hyvl	Enron Attorney
0.44	stephen.j.kean	VP and Chief of Staff
0.42	kevin.hyatt	Director-Pipeline Business
0.40	richard.b.sanders	VP Enron Services
0.39	elizabeth.sager	Asst. General Counsel
0.30	rick.buy@ enron.com	Manager - Risk Mgmt
0.30	mark.e.haedicke	MD - Legal Department

Table 2. Top 10 Propensities, Trained on NAS

same prevalence of privileged communication. We set aside 536 unlabeled families for testing. One way of evaluating the resulting propensity scores of the nodes in the test set is by examining the job roles (when known) of the people in the test set who are assigned the highest propensity scores. Tables 1 and 2 show the top 10 privileged persons from our person-scoring algorithm when trained on AS or NAS, respectively. The lists obtained by training on AS and NAS quite clearly find different people (Mark Taylor and Gary Taylor are the only common persons among the top 10). Based on role information from two sources (McCallum, 2005 and Shetty, 2005), from email signature blocks, and from other Web resources, we note several people on each list whose roles comport well with our intuition (lawyers, other legal staff, and senior executives). Our results reinforce our expectation that privilege can arise in part from the participation of a legal professional in the communication. Moreover, the results suggest that applying our technique to email from other organizations (for which ground truth role labels are not as easily available as for the widely studied Enron case) could indeed be useful.

Conclusion & Future Work

Our broad interest in supporting search amidst sensitive content has motivated our work, and the growing concern over the cost of privilege review in e-discovery has provided a setting for the research reported in this paper. The principal focus of this work has been on developing one way of automatically computing useful parameters that determine privilege, using both power iteration and (for people not seen in training) heuristic expansion through ego-centric networks. Our initial evaluation, based on inspection of the roles of highly ranked people, suggests that our method is able to identify several people who we would expect to engage in privileged communication. This is work in progress and we plan to expand it along several dimensions. We plan to utilize the computed propensity scores in an interactive privilege review system. For that, an obvious next step would be to explore techniques to score not only people but also content.

Looking to the longer term, what we learn through our experience with e-discovery can also inform our thinking about techniques that could be applied in other settings where search among sensitive content is at issue, including applications involving medical records, government transparency, and personal life-logging.

ACKNOWLEDGMENTS

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

REFERENCES

Boldi, P., et al. (2005). PageRank as a Function of Damping Factor. In WWW.

Cormack, G.V., et al. (2010). Overview of TREC 2010 Legal Track. In TREC.

Epstein, E. S. (2001). The Attorney-Client Privilege and the Work Product Doctrine. American Bar Association.

McCallum, A., et al. (2005). The Author-Recipient-Topic Model for Topic & Role Discovery in Social Networks, with Application to Enron & Academic email. LACS.

Shetty, J., et al. (2005). Discovering Important Nodes through Graph Entropy, the case of Enron email Database. In KDD Workshop.

Vinjumur, J. K. (2015) Evaluating Expertise & Sample Bias Effects Privilege for Classification in E-Discovery ICAIL

Xing, W.,et.al(2004)Weighted PageRank Algorithm. CNSR