Supervision and Discovery of Electronic Communications in the Financial Services Industry

Stefan Edlund, Tyrone Grandison, Joshua Hui, Christopher Johnson*,

IBM Almaden Research Center, 650 Harry Rd, San Jose, CA 95120
{sedlund, tyroneg, jhui}@us.ibm.com, chrisjohnson@mba.berkeley.edu

Abstract. Current SEC and NASD rules require securities brokers and dealers to maintain, supervise, and periodically review electronic communications. We present a solution called Galaxy that provides automatic supervision and in-depth discovery of email, instant messages, and other electronic communications to enable compliance with these rules. Galaxy's supervision component analyzes these communications to enforce company policies and detect potential violations. It allows compliance officers to generate powerful and flexible rules to implement information screens within an organization and detect suspicious text patterns in incoming and outgoing communications. Galaxy's discovery component enables companies to respond to litigation discovery requests efficiently. It also supports internal investigations by allowing analysts to focus their results along various search dimensions and visualize relationships among entities. In this paper, we describe Galaxy's architecture, illustrate the functionality of its supervision and discovery components using financial services scenarios, and propose topics for future research.

Keywords: Compliance, Supervision, Discovery.

1 Introduction

The United States Securities and Exchange Commission (SEC) and National Association of Securities Dealers (NASD) require securities brokers and dealers to maintain and supervise incoming and outgoing communications to ensure compliance with federal securities laws. These rules require improved technologies to monitor and search electronic correspondence. SEC Rule 17a-4 [1] requires exchange members, brokers, and dealers to maintain all email and other communications sent or received, including all inter-office memoranda and other communications, for a period of three years. NASD Rule 3010 [2] requires its members, which include brokers and dealers participating in the over-the-counter securities market, to establish and enforce procedures to supervise incoming and outgoing written and electronic correspondence. Rule 3010 also requires members to conduct periodic reviews of their business activities to assist in promoting compliance with, and detecting violations of, applicable securities laws and regulations. SEC Rule 10b-5 [3] and supporting case law prohibit companies and individuals from trading on inside

* Work was done while author was at IBM.

information or otherwise engaging in fraud or deceit in the purchase or sale of securities.

Galaxy is a communication management system that enables: 1) the supervision and discovery of electronic communications to facilitate compliance with SEC and NASD rules, and 2) timely and efficient responses to litigation discovery requests. Galaxy leverages prior research in the field on multi-faceted search [4] and text analytics [5]. Given that most forms of electronic communication contain a high proportion of free-form text, the Galaxy solution must: 1) detect and resolve errors, abbreviations, and acronyms, 2) provide an acceptable balance between false positives (precision) and false negatives (recall ratio) for compliance violation alerts, and 3) minimize the performance impact of the technology on daily business functions. Galaxy's analytic capabilities allow companies to intercept suspicious electronic communications in transit, detect suspicious text patterns in archived communications that may indicate violations of securities laws, and reduce the time and cost necessary to comply with the litigation discovery requests.

The application of a general system for discovery and supervision to real problems in a specific industry demonstrates the value of domain-focused solutions. In section 2, we define key terms necessary for our discussion on compliance in the financial services industry. In section 3, we describe the architecture of the Galaxy technology. We present the supervision and discovery components of Galaxy in sections 4 and 5, respectively. Finally, we discuss related work in section 6, future work in section 7, and conclusions in section 8.

2 The Environment

As this instantiation of Galaxy is intended for the financial services industry, we explain a few foundational terms and concepts before proceeding with the technology discussion. Specifically, we define the following example roles for managing and monitoring corporate communications, and the responsibilities of each role. We refer to these roles in describing the Galaxy technology and application scenarios.

- *Compliance officer:* At the direction of senior management, this role is responsible to implement policies and procedures, such as information screens[1] to supervise electronic communications in compliance with applicable securities laws. The purpose of an information screen is to monitor certain kinds of communication between people or groups and to block any communications that violate company policies and procedures. Because this is a sensitive role that frequently handles confidential internal information, companies may designate multiple officers, each responsible to supervise a subset of communications. Thus, it is desirable to control access to certain information about monitoring and supervision.
- *Internal auditor:* When a communication is flagged and intercepted for potential violation of corporate policy, an internal auditor receives it for further review. The auditor can either take no action if the communication complies with policy,

[1] Information screens are mechanisms that prevent information in an organizational silo from being disseminated in violation of company policies and procedures.

or flag a communication for further review. When a violation such as insider trading is suspected, the communication is forwarded to an internal investigator.

- *Internal investigator:* This role handles internal investigations, which may be triggered by an internal auditor, a suspicious company officer, an employee complaint, or a request from regulatory agency. An internal investigator will gather all evidence relevant to a case by extracting electronic information from company archives, such as past email communications, instant messages, trading records, and telephone records.
- *Discovery coordinator:* This role handles all legal discovery requests, including collecting all the requested documents (or other electronic evidence) and preparing a report summarizing all the collected information.

In small companies, the role of compliance auditor, internal investigator and discovery coordinator can belong to the same person. Having defined these roles, we describe the Galaxy technology and several application scenarios in the following sections.

3 Technology Overview

Galaxy assists companies in complying with regulations requiring supervision of electronic communications. Thus, the initial objective of the system is to reduce the amount of suspicious communications that are allowed by current systems. Galaxy is intended to detect clear violations of policies, mistaken disclosures, and indicators of improprieties that lead to deeper investigations. However, it is not intended to detect highly sophisticated violations, such as disseminating insider information using secret codes, or those violations perpetrated outside of the electronic communication system.

Because technology alone cannot detect all violations of company policies or securities laws, Galaxy must be administered by a compliance team to ensure that automated policies and procedures reflect ever-changing corporate and regulatory environments. This team should also be able to investigate potential violations detected by the automated system.

Figure 1 below shows an overview of Galaxy's architecture when it is integrated into an email archiving system. Each component represents an annotator that appends an additional piece of structured information to the email. For more information about annotators refer to the Unstructured Information Management Architecture (UIMA) documentation [6].

The initial annotator, i.e. the Meta Extractor in Figure 1, is fed a Multipurpose Internet Mail Extensions (MIME) encoded email document and extracts metadata, such as senders, recipients, subject and date information. The Group Extractor annotator then retrieves group information about sender and recipients by accessing an identity repository, such as a corporate directory. Ideally, corporate directories are cached locally to improve performance. Access to real-time corporate directory data is not essential; a cache that is updated daily or even weekly would be acceptable.

Galaxy then passes the communication to the Screen Extractor annotator, which extracts all information screens that the email crosses by consulting a policy database. A local cache helps improve performance by avoiding direct access to the policy

database. If there is no information screen involved, most of the subsequent annotators can be skipped except for tokenization/indexing.

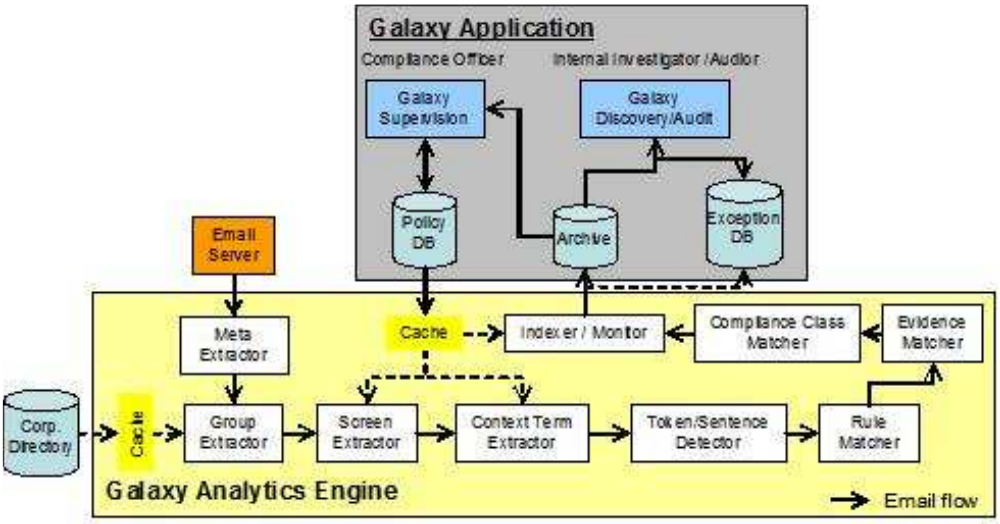

Fig. 1. Galaxy architecture integrated into an email archiving system.

If the Screen Extractor finds applicable information screens, the communication is tokenized and sentences are detected. Next, the Context Term Extractor retrieves any additional background information about the sender and incorporates this into the metadata. This may be a specific list of sensitive keywords that depends on the sender's current role or project. Such keywords are incorporated into the text analysis by extending Galaxy's native dictionary. Finally, we apply the three last annotators: the Rule Matcher performs semantic rule matching against the email, the Evidence Matcher combines matched rules into evidence categories, and the Compliance Class Matcher calculates compliance class scores. The Indexer/Monitor component is a UIMA [6] consumer with two responsibilities. First, it inserts the document into an email archive and text index so it can be searched by the Galaxy discovery component. Secondly, it determines whether the compliance class scores are above a threshold indicating the email should be reviewed by an internal auditor. If so, it creates a record in the exception database. Compliance officers can set thresholds and record them in the policy database.

The output from any of the annotators can be preserved and recorded in an email archive. For instance, it may be useful to preserve group or departmental information since company directories may not maintain historical information and employees tend to move around within companies. Galaxy also preserves supervision scores so a discovery coordinator can organize and sort emails accordingly. Currently, we do not store information about evidence categories found in the Galaxy discovery archive, but this is an easy extension that would allow discovery coordinators to ask questions such as "Which emails discuss buying or selling of stock in company A?," assuming such an evidence category has been built.

We describe each of these processes and components in more detail in the context of the application scenarios.

4 Supervision

In this section, we describe three example scenarios to illustrate the features and benefits of the Galaxy supervision system in an investment banking environment. The first scenario involves supervising communication among various departments to detect inappropriate disclosures of insider information. The second scenario concerns supervising communications between employees and outsiders to detect insider information leakage. The third involves screening employees from communications that would create conflicts of interest.

4.1 Supervision Scenarios

BankCo is an investment bank regulated by the SEC and NASD that is implementing procedures to supervise electronic communications. It would like to impose information screens to detect improper flows of insider information among departments.

Screening Among Departments. BankCo wants to supervise communications between the equity research department and the brokerage department to ensure that brokers do not have knowledge of stock ratings (e.g. buy/sell/hold) before they are disclosed to the market. These ratings often cause movement in the stock price and BankCo would like to assure that its brokers and their customers are not able to profit from this information in violation of the securities laws. BankCo also wants to supervise communications between the mergers and acquisitions department and the brokerage department to assure that its brokers and clients do not have access to insider information about pending acquisitions. Thus, it wants to create email supervision rules that automatically detect suspicious emails and intercept them for review.

Screening from Outside Entities. Communication between BankCo employees and outsiders concerning insider information about pending transactions, such as stock buybacks, acquisitions, and planned purchases of large blocks of securities must be supervised. Of course, BankCo works with a number of outside companies that are privy to this information. Thus, there is a challenge in reducing false positives and intercepting only communications that pass insider information to unauthorized parties.

Screening Conflicts of Interest. Another supervision scenario involves screening information from particular employees to avoid potential conflicts of interest, so called “Chinese Walls”. Suppose that BankCo hires an associate that previously worked in the corporate finance department of TechnologyCo. Shortly thereafter, BankCo represents ParentCo in the acquisition of TechnologyCo. The new associate is legally and ethically bound not to disclose to BankCo any of the insider information that he obtained while working at TechnologyCo. BankCo must screen this employee from any communications involving the acquisition deal. Therefore, it would like to

intercept any communications to or from this employee that refers to TechnologyCo or the pending deal. Clearly, BankCo needs flexible supervision technology that adapts to a broad range of scenarios.

4.2 Galaxy's Supervision System

The ability to construct and dismantle electronic information screens is an essential component for supervision. Galaxy enables a compliance officer to access and view company information screens, but only those screens originally created by the officer. Also, a compliance officer cannot typically define screens that monitor traffic sent or received by him, since this creates a security risk.

Creating Information Screens

In Galaxy's supervision system, an information screen has the following components:

- **Sender ID**. This is either the name of an individual or a group, such as departmental information. The information screen is only in effect if the sender matches this attribute.
- **Receiver ID**. This has a similar definition to sender ID, but the attribute is matched against the receiver side of a communication. The screen is applied only if at least one of the recipients, by name or associated group, matches this attribute.
- **Compliance classes**. A compliance class monitors communication to intercept one particular kind of violation, e.g. insider trading. We define in detail below what constitutes a compliance class.
- **Start/End date**. This is an optional attribute that allows an information screen to be activated or deactivated automatically on the dates specified. This is useful, for instance, if a small team is working on a sensitive project for a pre-determined period of time, e.g. an acquisition.

We augment the system by allowing special tags for internal and external roles in the sender/receiver attributes. This allows specification of information screens that monitor all incoming and outgoing traffic. Roles may also be defined for other external roles such as attorneys and accountants. As an integrity check, all tasks carried out by compliance officers produce an audit record in a log file.

Compliance Classes

We define a *supervised email* as an email associated with one or more compliance class scores. As defined above, a compliance class is part of a company policy and represents one type of suspicious communication being monitored. An information screen is associated with one or multiple compliance classes. When an email is sent, it is possible that more than one information screen is crossed. For instance, a memo sent from the equity research department to the broker division of an investment bank, as well as to an external address, crosses two information screens. Each screen is associated with a set of compliance classes, which may overlap one another. As part of the analytics, we collect the unique set of compliance classes associated with the

email, $\{C_1, \ldots, C_n\}$. For each compliance class, a score $score(C_i)$ is calculated, such that $0 \leq score(C_i) \leq 1$, where $0 \leq i \leq n$.

Evidence Categories and Semantic Rules

Each compliance class is created by combining **items of evidence** found in the email. An **evidence category** E is a collection of **semantic rules** (defined below) that describe a common concept or intent. For example, suppose E_m is an evidence category that states that part of the email is discussing a private meeting, E_{bs} is an evidence category that states that the email is discussing buying or selling of securities, and E_{fn} is an evidence category that states that an equity researcher gives advance notice of a future stock recommendation to a broker. Clearly, E_{fn} violates the bank's policies, but E_m and E_{bs} are only potential violations when detected in the same email. In this example, a compliance class monitoring suspicious communication $C_{suspicious}$ is a combination of the three evidence categories in the following way:

$$C_{suspicious} = (E_m \text{ and } E_{bs}) \text{ or } E_{fn}$$

Note that an evidence category E is associated with one or more semantic rules. In Galaxy, a *semantic rule* is a sequence of basic or generalized terms that are matched against sentences detected in the email. A *generalized term* can be either a term matching any synonym of a given term, e.g. buy, purchase, acquire, or a hypernym. Another example is "security" which would match "stock," "bond," "note," etc. Syntactically, we write a sample semantic rule as:

[buy]<security>

This rule matches any synonym of "buy" followed by any mention of a security in a single sentence. The rules are based on the extraction patterns that have been successfully used by Riloff et al. [7]. To enable the application of semantic rules, Galaxy uses a native dictionary of synonyms and hypernyms to match rules against text sentences. Since emails frequently have errors and misspellings, Galaxy also supports fuzzy matching of terms based on the Levenshtein editing distance metric [8].

Calculating Scores

Each rule R has an associated weight, $weight(R)$, where $0 \leq weight(R) \leq 1$. The weight represents the "accuracy" of the rule. A rule with a high probability of matching the concept or intent being captured by the evidence category has a weight close to 1, while less accurate rules have weights closer to 0. Currently, we use a heuristic method in setting the weight depending on the confidence of the rule creator. We suggest that rules be validated against the email archive to evaluate the precision of the rule. In the future, the method for determining the associated rule weights will be formalized and improved. If a rule matches an email, we do not attempt to match the same rule against the email again. Instead, we say that:

$$match(R_i, email) = true \quad (\mathbf{1})$$

When calculating the evidence category score for an evidence category E, we select the score with the highest weight that also matches the email:

$$score(E) = MAX(weight(R_1), .., weight(R_n)) \quad (\mathbf{2})$$

where $R_i \in E$ and $match(R_i, email) = true$, $1 \leq i \leq n$

From the range restriction on weight(R_i), we also see that $0 \leq score(E) \leq 1$. The final step in this process is the computation of the score for the compliance class. Since a compliance class is built from a combination of evidence categories, we define how to combine such expressions:

$$score(E_1 \text{ or } E_2 \text{ or } ... \text{ or } E_n) = \quad (3)$$
$$MAX(score(E_1), score(E_2), \ldots, score(E_n))$$

$$score(E_1 \text{ and } E_2, ..., \text{ and } E_n) = \quad (4)$$
$$(score(E_1) + score(E_2) + ... + score(E_n)) / n$$
$$\text{iff } (score(E_x) > 0) \text{ for } 1 \leq x \leq n,$$
$$\text{otherwise } score(E_1 \text{ and } E_2, ..., \text{ and } E_n) = 0$$

Intuitively, the score for a compliance class is the maximum score of all matching evidence categories. When more than a single evidence category must be detected in the email, we average the score of the matching evidence categories. As a result, the final score for a compliance class is between 0 and 1.

Internally, when email text is analyzed, we optimize the text matching by combining all relevant semantic rules into a single **state machine**. The state machine is passed one sentence at a time and when the sentence is completely processed, all states are reset. However, when a final state is reached, i.e. a rule is matched, we can further optimize the rule processing by removing that rule from the state machine, as we need not match the same rule again against the same document. This ensures that processing overhead is reduced and overall throughput is improved, a critical requirement for email supervision systems.

4.3 Administration and audit tooling

A supervised email with associated compliance class scores can be handled several different ways by an application. Galaxy produces a "Work Item" list from the exception database where a internal auditor can access and sort emails according to either an aggregate compliance class score, or narrow down to specific compliance classes. Each flagged email can be retrieved and parts of the email that match semantic rules are highlighted. When an auditor moves the mouse over the highlighted text, the syntax of the matching rule is displayed. This provides feedback to indicate the reason why the email was flagged.

Galaxy also provides a method of incorporating new semantic rules. It allows users to enter a sentence directly, where the sentence is analyzed and a suggested semantic rule is returned. Words that are recognized, from Galaxy's dictionary, can be generalized or specialized as desired. Galaxy enables the user to validate the rule against an email archive to see whether it returns meaningful documents. This improves the precision by reducing false-positives.

When the user is satisfied with a new rule, he can associate the rule with an existing evidence category and compliance class. Semantic rules can be reused in many evidence categories, and evidence categories can be used in more than one compliance class.

Finally, Galaxy provides a dictionary tool that uses Latent Semantic Analysis (LSA) [9], which is useful for finding synonyms of a given term or phrase. LSA is an

unsupervised machine learning method that is completely language independent. However, LSA can be resource intensive, so a sample of business communication is extracted before analysis. The sampling can either be random or targeted towards a particular business function, such as client/broker communication. A compliance officer can use LSA to add new rules by finding other ways employees express terms or phrases. This can improve recall ratio by reducing the overall false-negatives detected by the supervision function.

5 Discovery

For Galaxy's discovery component, we provide two example scenarios. The first scenario demonstrates how Galaxy can be used to respond to electronic discovery requests in the course of litigation. The second scenario shows the unique features of Galaxy that facilitate internal investigations of company archives. For each of these scenarios, we describe the capabilities and advantages of the Galaxy system.

5.1 Litigation Discovery

Litigation discovery is the process whereby the parties to a lawsuit request and exchange documents and other material evidence. The discovery process often involves the exchange of electronic communications, such as archived email or instant messages. Given the volume of many company email archives, responding to these requests can be particularly burdensome, time-consuming, and expensive. Consider the following scenario:

BankCo operates many large mutual funds. A group of mutual fund investors sues BankCo in a class action lawsuit, alleging that BankCo has improperly favored select customers by executing their trades in advance of the mutual fund trades. They claim that this practice, known as "front running," has resulted in significant lost returns to mutual fund investors. By executing individual trades prior to large block mutual fund trades, the brokers allegedly allowed select individuals to profit by the subsequent share price increases, at the expense of mutual fund returns. After filing their lawsuit, the plaintiff class serves BankCo with written discovery requests, including 35 requests for production of documents. These requests seek various documents and electronically stored information relevant to the causes of action asserted in the lawsuit. The following are two examples of plaintiffs' discovery requests:

1. All email and other electronic communications between BankCo and customers Arthur, Barbara, and Carter concerning companies TechnologyCo, PharmCo, and FoodCo between January 1, 2004 and present.
2. All email and other electronic communications sent or received by BankCo brokers regarding mutual funds FutureEnergy, NextStep, and ValueLife, also having connection with customers Arthur, Barbara, and Carter, between January 1, 2004 and present.

Responding to these 35 discovery requests requires BankCo to review all the electronic communications to identify all responsive email, instant messages, and electronic documents. However, the accuracy of this process is very important, as the parties are required to produce all non-privileged evidence that is responsive to the discovery requests. They must also make sure to remove any privileged documents, such as attorney-client confidential communications and attorney work product, and also any documents protected by constitutional or statutory privacy rights, prior to providing responsive documents to the other party. BankCo would like to have an accurate and reliable method to conduct this review so it can save the considerable time and expense of a manual review.

To handle the first request, we must first identify all the relevant terms (or synonyms) corresponding to TechnologyCo, PharmCo, and FoodCo, which may include the company names (full or short), the stock symbols, and any other potential nicknames (e.g. "Big Blue" would refer to IBM). Other search conditions include the date range (e.g. from January 1st, 2004 to now), the sender and recipient list (the three customers) and a list of terms regarding company TechnologyCo, PharmCo, and FoodCo, against the body and attachment. As previously mentioned, the responding party needs to filter out any privileged documents. Therefore, we add all of the lawyers' email addresses as negation terms to the sender and recipient list. However, this may not guarantee that we have the complete list. Manual review, usually by the legal team, is required to further verify and sanitize the documents to be produced.

The second request is similar to the first one, which also includes any search term corresponding to the three mutual funds, and the specified date range. However, it is not obvious to determine the senders and recipients list. Naively, we can put no constraint in the address list. But the result, R, will include all the brokers who have mentioned and processed any trade request for those three mutual funds, regardless of any connection to the three customers. In order to better estimate the contact list, we first generate a social network graph, G(V, E), from the result, R. V is a set of vertices, one for each email address from a sender, v_i, and a recipient, v_j, of R, and E is a set of edges, $\{e^{v_i\_v_j}\}$, if there is a communication from v_i to v_j. Note that each edge is an aggregation of all communication between sender v_i and recipient v_j, rather than a single communication. Assuming S is a set of brokers which we have identified as part of the first discovery request (i.e. brokers interfacing with Arthur, Barbara and Carter), we can estimate the broker list by computing the reachable vertices from S. Since we assume the information should be flowing from any broker handling mutual fund transactions to the brokers, S, we traverse the graph in the reverse direction. Figure 2 describes the algorithm. The adjacentVertices() function returns the adjacent vertices of all the inbound edges from a given vertex.

```
Reachability(S) {
    CurrV <- S, TotalV <- S
    do {
       NewV <- {}
       for ( v in CurrV ) {
          AdjV <- adjacentVertices(v)
          for ( a in AdjV )
               if ( a is not in TotalV )
                   NewV <- NewV + a
```

```
    }
    CurrV <- NewV
    TotalV <- TotalV + NewV
} while (NewV is not empty)
```

Fig. 2. Reachability Algorithm

We can further reduce the broker list if the event sequence is also considered. For each edge, $e^{v_i\_v_j}$, we store an attribute, [first($e^{v_i\_v_j}$), last($e^{v_i\_v_j}$)], which contains the date of first communication, first($e^{v_i\_v_j}$), and the date of the last communication, last($e^{v_i\_v_j}$), from user v_i to v_j from the result, R. Now we define a vertex, v_k, as being reachable from v_j if and only if either v_j belongs to the initial vertices set, S, or last($e^{v_j\_v_k}$) is equal or later than the earliest start date, min(first($e^{v_i\_v_{i+1}}$)), of any reachable path, (v_1,..,v_i,v_{i+1},..,v_j), from the initial set, S. To compute the broker list, we change the adjacentVertices() function to return the next set of vertices only if they satisfy this condition.

By further reducing the search result, Galaxy lessens the amount of manual review needed to produce the requested documents.

5.2 Internal Investigation Discovery

An internal investigation discovery request is usually triggered by an employee complaint or a suspicious event identified in the supervision process. Therefore, it usually has some initial starting points, such as the target subject, or some involved parties. However, there are still a lot of unknown factors. So the main task of the internal investigator is to uncover the unknown.

However, the simple form-based search and plain result list interface (similar to web search) is not sufficient for this task. This method is useful for finding information such as restaurant reviews or product information, because as long as relevant results are returned in the first two pages, users do not look at the rest of the search results. For this reason, the ranking algorithm is very important. However, for internal investigation discovery, each email has the same level of importance, like pieces of a puzzle, in reconstructing a sequence of events. At the same time, there can be many hits in the search result and it is difficult and time-consuming to navigate and understand all of them. The internal investigator wants to narrow the search and filter out the irrelevant ones in a systematic manner. Consider the following internal investigation scenario:

BankCo would like to investigate its mergers and acquisitions (M&A) department to determine whether employees have improperly disclosed confidential client information. BankCo wants to investigate three specific client acquisitions. Because its electronic archives are too large to conduct this investigation manually, BankCo would like to search the archives for suspicious phrases and patterns, including all email, instant messages, and other communications sent by anyone in the mergers and acquisitions department, containing the name of clients: TechnologyCo, PharmCo, or FoodCo, between the time of the initial client meetings and the merger announcements.

When constructing the discovery request, the investigator must first identify the potential suspects. The initial set would include employees working in the M&A department, employees working in the brokerage department, any employee working in the client companies, and outside recipients of confidential information. However, if we use this list and the client's company name for the discovery search, the search result will be huge and will be unlikely to provide much useful information. Therefore, the search must be further refined as follows:

- Using text classification to filter out any unrelated document, such as meeting invitations or general announcements; and
- Using additional related search terms may include "merger," "acquisition," "stock purchase," "tender offer," plus their synonyms and hypernyms.

Upon paring the communications down to a suspicious set, we can examine the social networks of the senders and recipients of those communications, and their threads, to uncover any patterns or additional persons involved. In the next section, we describe how Galaxy enables this kind of analysis.

5.3 Supervision Scenarios

Galaxy's discovery component can logically divide it into two sections. The first section, called *Basic Search*, consists of the search bar, the result table, and the email preview panel. It is a typical search panel which can be found in many other legal discovery products. There are a number of things we have added to enhance the discovery functionality, including the violation score and the user profile lookup in the preview panel. The violation score is the value which we have computed during the supervision phase, as described in the section 4.2. The user profile not only includes the basic user information, such as job title, and the department, but also includes the aggregate violation score, which consists of all the previous flagged communications, using the following formula.

$$\sum_{e \in E} \left(score(e) * (exist\ (sender(e), u) + \frac{exist(recipient(e), u)}{|\ recipient(e)\ |^2}) \right)$$

where E is a set of all the flagged emails with score > 0, and exist is a function which checks for the existence of the given user, u, in the input list, and returns {0, 1}. Each score is weighted based on the sender and recipient list. We use the term $|\ recipient(e)\ |^2$ to model the diminishing effect that a large number of recipients has on the score. That is, the more people on the recipient list the less the score value that should be added to the accumulative score value. We also foresee that more compliance-related information can be added to the same user profile area to facilitate the discovery process further. This includes all previous and current client history and transactions by the selected person executed around specific days.

The second logical section, called *Summarization*, provides the multi-faceted search capability on the result set. This section contains different visual representations of information extracted from the result set. Currently, Galaxy provides visualizations based on summarizations of the top-N senders, top-N receivers, sent-date distributions, classifications and social networks. The list can be extended to include other visual techniques, such as Tag-Clouds [10] (which provide

visual representations of words based on the frequency) and stack graph, as seen in the Many Eyes project [11]. Users can further refine the search by drilling down (i.e. dragging and dropping the refined dimension into the drill-down basket) into the various dimensions such as *category* for classification chart, *people* for the social network and top-N charts, and *date range* for the date distribution. Regarding the social network, each user node contains the aggregate violation score (the same one extracted from the user profile), plus the score for each category. We can also eliminate any broadcast communication by examining the number of recipients. Furthermore, we provide different algorithms to highlight a specified set of users in the network, such as the reachability set (as described in section 5.1) and the top N clustered users based on the aggregate score.

Apart from the regular search-oriented discovery interface, Galaxy's discovery component also provides a temporal analyzer, which aligns multiple search results together with external events (e.g. stock price or trading history) on the same time dimension. This tool correlates events and annotates the search results.

6 Related Work

There are several email supervision products currently available, such as Orchestria [12], CA Message Manager (previously known as iLumin) [13] and Zantaz [14]. These products use linguistic pattern matching techniques to tokenize the document and search for suspicious patterns. Since there are many ways to express the same idea via electronic communications, and these ways vary among industries and regions, the rules must be continually tuned and updated to achieve high accuracy. However, for some products, like the CA Message Manager, the actual rules are hidden from the users (i.e., in a "black box"), and are therefore difficult to modify. These systems often require professional services from the vendor for hand-tuning. Other vendors allow customization, but rely on the user's linguistic knowledge to construct the precise regular expression or pattern. Galaxy takes a different approach, allowing users to derive rules from a sample sentence. Galaxy also categorizes the rules using higher level concepts (i.e., evidence categories), which are easier to maintain, especially if the rule base is large (i.e., several thousand rules).

On the discovery side, commercial products, such as Zantaz [14], Symantec Enterprise Vault [15] and ZipLip [16], provide typical text search interface, including fuzzy and proximity search capabilities, on the metadata fields such as *from*, *to*, *date* and *subject*, as well as on the body and the attachment. We refer to this as *Basic Search*. However, none of these systems provide the advanced multi-faceted search interface of Galaxy, which guides discovery coordinators and internal investigators to understand and further filter the search results.

Some research projects, like EMT [17][18], employ visualization and mining techniques to analyze and detect anomalies against the email data. This functionality can also be incorporated into the Galaxy discovery framework, if desired.

ADS [19] and SONAR [20] are fraud detection systems developed by NASD which use a variety of AI techniques, including visualization, pattern recognition, and data mining, in support of the activities of regulatory analysis, alert and pattern detection.

They focus on mining the transaction data, together with external data, such as news feeds, but they do not link this data with other unstructured information, such as electronic communications [21]. By analyzing electronic communications, Galaxy can provide additional context necessary to improve positive detection rates.

7 Future Work

Galaxy is an evolving system and is currently being tested and evaluated to determine the weaknesses and strengths of the system's assumptions and approach. One important area for future research is to develop supervision and discovery technologies that preserve individual privacy. For instance, a privacy-preserving discovery tool could obscure certain sensitive information in intercepted email communications. Similarly, improved discovery tools could incorporate technologies that de-identify unstructured text without significantly degrading the accuracy of the forensic analysis.

A second useful enhancement of Galaxy is improving the methods used to aggregate and summarize search results. When the number of search results is large, a method for sampling results can dramatically outperform an exhaustive enumeration. Since search results typically are not randomly ordered (more commonly results are ordered by a ranking algorithm), the sampling method and sample size must be carefully selected. Thus, we would like to evaluate and implement sampling algorithms that would allow Galaxy to summarize search results more efficiently.

A third area of research involves detecting patterns in electronic communication archives that may indicate policy violations or other improprieties. For example, assume that periodic communications from the same sender result in small trades based on insider information. However, this activity is not detected by the supervision system because the tips are non-obvious and the resulting trades, in isolation, are not considered material. If this communication pattern could be traced and the trades were considered in the aggregate, the financial impact would be significant. Therefore, further research is necessary to detect such patterns in email archives.

8 Conclusion

Compliance with SEC and NASD rules is a critical requirement for information systems used by financial services companies. In this paper, we introduced Galaxy, which provides advanced supervision and discovery capabilities for various forms of electronic communication. We also discussed the advantages that Galaxy offers over other available technologies. Galaxy empowers companies to enforce supervision policies and procedures, search data archives, and conduct internal investigations. It is an extensible solution that improves supervision of electronic communications by leveraging UIMA technology, an error tolerant and scalable pattern matching engine and latent semantic analysis. Further, Galaxy improves the efficiency and quality searching electronic archives to respond to litigation discovery requests. Although this paper describes application scenarios from the financial services domain, Galaxy can

be applied to any industry requiring similar supervision or discovery capabilities. We hope that this work will be useful to the research community in developing more convenient and useful methods of managing electronic communications.

References

1. SEC Rule 17a-4, http://www.sec.gov/rules/final.shtml
2. NASD Rule 3010, http://nasd.complinet.com/nasd/display/index.html
3. SEC Rule 10b-5, , http://www.sec.gov/rules/final.shtml
4. Hearst, M. “Next Generation Web Search: Setting Our Sites,” IEEE Data Eng. Bull. 23(3): 38-48 (2000)
5. Weiss, S.M., Indurkhya, N. , Zhang, T. and Damerau, F. "Text Mining: Predictive Methods for Analyzing Unstructured Information," Springer, March 2007, 236 pages, ISBN 0387954333
6. Unstructured Information Management Architecture (UIMA), http://incubator.apache.org/uima/
7. Riloff, E., and Jones, R. "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99) , 1999, pp. 474-479
8. Levenshtein, I. V. “Binary codes capable of correcting deletions, insertions, and reversals,” Doklady Akademii Nauk SSSR, 163(4):845-848, 1965.
9. Landauer, T., Foltz, P.W., and Laham, D. "Introduction to Latent Semantic Analysis". Discourse Processes 25: 259-284. 1998
10. Hassan-Montero, Y., and Herrero-Solana, V., "Improving tag-clouds as visual information retrieval interfaces," Proc. InfoSciT 2006
11. Many Eyes, http://services.alphaworks.ibm.com/manyeyes/home
12. Orchestria, http://www.orchestria.com/
13. CA Message Manager, http://www.ca.com/us/products/product.aspx?ID=5707
14. Zantaz, http://www.zantaz.com/
15. Symantec Enterprise Vault, http://www.symantec.com/enterprise/products/overview.jsp?pcid=1018&pvid=322_1
16. ZipLip, http://www.ziplip.com/
17. Li, W., Hershkop, S., and Stolfo, S. “Email Archive Analysis Through Graphical Visualization,” Proc. 2004 ACM workshop on Visualization and data mining for computer security, Washington DC, USA, 2004
18. Stolfo, S., Creamer, G., and Hershkop, S. “A Temporal Based Forensic Analysis of Electronic Communication,” 2006, Digital Government Proceedings, San Diego, CA
19. Kirkland, D., Senator, T., Hayden, J., Dybala, T., Goldberg, H. and Shyr, P. “The NASD Regulation Advanced Detection System”. AAAI 20(1): Spring, 55-67.
20. Goldberg, H., Kirkland, J., Lee, D., Shyr, P. and Thakker, D. “The NASD Securities Observation, News Analysis and Regulation System (SONAR)”. Proc. of IAAI03
21. Phua, C., Lee, V., Smith, K., and Gayler, R. “A comprehensive survey of data mining-based fraud detection research,” Artificial Intelligence Review, 2005.