

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 Road, San Jose, CA 95120
{sedlund, tyroneg, jhui}@us.ibm.com

Abstract. We present a system called Galaxy that assists financial services companies in supervising electronic communications, searching electronic information archives to conduct internal investigations, and responding to legal discovery requests. Galaxy's supervision component enables companies to implement electronic information screens, identify suspicious text patterns in electronic communications, and detect and prevent violations of policies and procedures, in compliance with applicable securities laws and self-regulatory organization rules. Galaxy's discovery component supports in-depth internal investigations by allowing companies to search electronic information archives, focus their results along various search dimensions, and visualize relationships among entities. It also enables companies to compile responses to legal discovery requests for electronic information more efficiently. In this paper, we describe the architecture of Galaxy's supervision and discovery components, illustrate the functionality these components using example scenarios from the financial services industry, and propose topics for future research.

Keywords: Compliance, Supervision, Discovery

1 Introduction

United States federal securities laws and self-regulatory organization (SRO) rules require regulated financial services companies, including registered securities brokers and dealers, to establish systems and procedures to supervise electronic communications. National Association of Securities Dealers (NASD) Rule 3010 [1] requires each of its members to establish and maintain a system to supervise the activities of each registered representative, registered principal, and other associated person that is reasonably designed to achieve compliance with applicable securities laws and regulations, and with NASD rules. New York Stock Exchange (NYSE) Rule 342 [2] also requires its members to implement supervisory policies and procedures. Financial Industry Regulatory Authority (FINRA)¹ members must have policies and procedures for the review by a supervisor of employees' incoming, outgoing, and

¹ FINRA was created in July 2007 through the consolidation of NASD and NYSE Member Regulation. FINRA is working toward a consolidated rulebook, but its current rulebook consists of NASD rules and certain NYSE rules that FINRA has incorporated.

* Work was done while author was at IBM.

internal electronic communications that require review under FINRA rules and federal securities laws [3]. In addition, Section 15(f) of the Securities Exchange Act of 1934 [4] requires registered brokers and dealers to establish, maintain, and enforce written policies and procedures reasonably designed to prevent the misuse in violation of the Act, or the rules or regulations thereunder, of material, nonpublic information by such broker or dealer or any person associated with such broker or dealer. Securities and Exchange Commission (SEC) Rule 10b-5 [5] and supporting case law prohibit insider trading or otherwise engaging in fraud or deceit in connection with the purchase or sale of securities. Accordingly, companies in the financial services industry require effective technologies for monitoring, searching, and analyzing electronic communications.

We propose a communication management system called Galaxy that: 1) provides supervision and discovery of electronic communications to facilitate compliance with applicable securities laws, regulations, and SRO rules², and 2) assists companies in searching electronic archives to compile responses to legal discovery requests. Galaxy leverages prior research in the field on multi-faceted search [6] and text analytics [7]. 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 supervise and intercept suspicious electronic communications, detect suspicious text patterns in archived communications, and reduce the time and expense of complying with legal discovery requests for electronic information.

In the sections that follow, we demonstrate a general system for supervision and discovery by describing its applications to real problems in a specific industry. In section 2, we define terms necessary for our discussion of the Galaxy system in the financial services industry. In section 3, we describe the architecture of 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 and responsibilities, which we refer to in describing Galaxy technology and several application scenarios.

- *Compliance officer:* At the direction of senior management, this role is responsible to implement policies and procedures, such as information screens, to supervise electronic communications in compliance with applicable securities laws and SRO rules. Information screens are mechanisms that prevent

² For example, FINRA, NASD, and NYSE rules.

information from being disseminated in violation of specified organizational policies. Compliance officers are also responsible to review certain suspicious communications. Because this is a sensitive role that frequently handles confidential internal information, companies may designate multiple compliance officers, each responsible to supervise a subset of communications.

- *Internal auditor:* When a communication is flagged and/or intercepted for potential violation of an organizational policy, it may be forwarded to an internal auditor for further review. The auditor can either: 1) take no action if he determines the communication complies with policy, 2) mark the communication for further review, or 3) forward the communication to an internal investigator if he suspects a policy violation.
- *Internal investigator:* This role handles internal investigations, which may be triggered by an internal auditor, a suspicious supervisor, an employee complaint, a request from a regulatory agency, or other means. An internal investigator will gather evidence relevant to an investigation by extracting electronic information from company archives, such as email communications, instant messages, trading records, and telephone records.
- *Discovery coordinator:* This role handles legal discovery requests, including compiling requested documents and/or other electronic information and preparing reports summarizing the collected information.

Having defined these example roles, we describe Galaxy technology and several application scenarios in the following sections.

3 Technology Overview

Galaxy assists companies in complying with federal securities laws and SRO rules requiring supervision of electronic communications. 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 and indicators of improprieties that may lead to further investigation. However, it may not detect highly sophisticated violations, such as disseminating inside information using secret codes or those violations perpetrated outside of the electronic communication system. Galaxy must be administered by a compliance team to ensure that its policies and procedures reflect ever-changing organizational and regulatory environments. Because technology alone cannot detect all violations of company policies or securities laws, the compliance team should 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 [8].

The initial annotator, 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

since company directories may not maintain historical information and employees tend to move around within companies. Galaxy also preserves supervision scores so an internal investigator 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 internal investigators to ask questions such as “Which emails discuss buying or selling of stock in company A?,” assuming that 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 a financial services company. The first scenario involves supervising communication among various departments to detect inappropriate disclosures of material, nonpublic information. The second scenario concerns supervising communications between employees and outsiders to detect improper disclosures. The third involves screening individual employees from communications that may create conflicts of interest.

4.1 Supervision Scenarios

BankCo is an investment bank that is required to establish a system and procedures to supervise electronic communications. Among other things, it would like to implement information screens to detect and prevent improper disclosure of material, nonpublic information among employees, departments, and outside entities. Below are a few examples of such information screens.

Screening Among Departments. BankCo wants to supervise communications between its equity research department and its brokerage department to ensure that brokers do not have knowledge of analyst reports or ratings (e.g., buy/sell/hold) before they are disclosed to the market. These ratings often cause movement in the stock price and BankCo wants to assure that its brokers and their customers are not able to trade on this material, nonpublic information in violation of the securities laws. BankCo would also like to screen electronic communications between its equity research department and investment banking department to prevent undue influence on research personnel, in compliance with SRO rules.

In addition, BankCo would like to supervise communications between its mergers and acquisitions department and its brokerage department to assure that brokers and their clients do not have access to material, nonpublic information about pending acquisitions.

Thus, BankCo wants to create and enforce supervision rules that automatically detect suspicious electronic communications and intercept them for review.

Screening from Outside Entities. BankCo would also like to supervise communications between its employees and outsiders concerning material, nonpublic

information about pending client transactions, such as acquisitions, stock buybacks, and planned purchases of large blocks of securities. Of course, BankCo works with outside companies that are also privy to this information. Accordingly, there is a challenge in reducing false positives and intercepting only communications that disclose material, nonpublic information to unauthorized parties.

Screening Individuals. BankCo would also like the ability to screen information from particular employees to avoid conflicts of interest. Suppose that BankCo hires an employee that previously worked in the finance department of TechnologyCo. Shortly thereafter, BankCo advises ParentCo concerning its potential acquisition of TechnologyCo. BankCo would like to prevent its new employee from disclosing any confidential information he may have obtained while working at TechnologyCo or improperly disclosing any information concerning the potential acquisition. Therefore, BankCo would like to screen this employee from sending or receiving any communications involving ParentCo, TechnologyCo, or the potential acquisition.

Clearly, BankCo needs flexible supervision technology that adapts to a broad range of scenarios.

4.2 Galaxy's Supervision System

An essential component of Galaxy's supervision system is the ability to implement and manage electronic information screens. Galaxy enables a compliance officer to access and view only those information screens that the officer originally created or is otherwise authorized to view. Also, it does not allow a compliance officer to define screens that monitor communications sent or received by him, as this would create a security risk. As an integrity check, all tasks carried out by compliance officers produce an audit record in a log file.

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 one or more recipients, by name or associated group, match this attribute.
- **Compliance classes.** A compliance class monitors communication to detect one particular type of violation. We define compliance classes in detail below.
- **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, such as a potential acquisition, for a limited period of time.

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 external individuals such as attorneys and accountants.

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 determined by 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 brokerage department 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 analysis, we collect the unique set of compliance classes associated with the email, $\{C_1, \dots, C_n\}$. For each compliance class, a score $\text{score}(C_i)$ is calculated, such that $0 \leq \text{score}(C_i) \leq 1$, 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 research analyst gives advance notice of a future stock rating 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 $C_{\text{suspicious}}$ is a combination of the three evidence categories in the following way:

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

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. [9]. 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 contain errors and misspellings, Galaxy also

supports fuzzy matching of terms based on the Levenshtein editing distance metric [10].

Calculating Scores

Each rule R has an associated weight, $\text{weight}(R)$, where $0 \leq \text{weight}(R) \leq 1$. The weight represents the “accuracy” of the rule. A rule with a high probability of matching the concept represented 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:

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

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

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

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

From the range restriction on $\text{weight}(R_i)$, we also see that $0 \leq \text{score}(E) \leq 1$. 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:

$$\text{score}(E_1 \text{ or } E_2 \text{ or } \dots \text{ or } E_n) = \quad (3)$$

$$\text{MAX}(\text{score}(E_1), \text{score}(E_2), \dots, \text{score}(E_n))$$

$$\text{score}(E_1 \text{ and } E_2, \dots, \text{and } E_n) = \quad (4)$$

$$(\text{score}(E_1) + \text{score}(E_2) + \dots + \text{score}(E_n)) / n$$

$$\text{iff } (\text{score}(E_x) > 0) \text{ for } 1 \leq x \leq n,$$

$$\text{otherwise } \text{score}(E_1 \text{ and } E_2, \dots, \text{and } E_n) = 0$$

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) [11], 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 assist companies in responding to electronic discovery requests in the course of litigation. The second scenario shows the features of Galaxy that facilitate internal investigations of company archives.

5.1 Legal Discovery

Legal 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 example scenario.

BankCo is a defendant in a class action lawsuit, brought on behalf of investors in three mutual funds, alleging that BankCo tipped select hedge fund customers prior to executing large block trades for the mutual funds, in violation of federal securities laws. Plaintiffs claim that these tips allowed the hedge funds to execute trades in the same securities ahead of the mutual fund trades, resulting in higher trading prices for the mutual funds (in the case of purchases) and lower returns for their investors. Plaintiffs served BankCo with a request for production of documents seeking various documents and electronically stored information relevant to the causes of action asserted in the lawsuit. The following are two example categories of documents requested by plaintiffs:

1. All email and other electronic communications between BankCo and customers HedgeA, HedgeB, or HedgeC concerning companies TechnologyCo, FoodCo, or PharmaCo between January 1, 2005 and present.
2. All email and other electronic communications sent or received by BankCo brokers regarding mutual funds MutualX, MutualY, or MutualZ, and concerning customers HedgeA, HedgeB, or HedgeC, between January 1, 2005 and present.

Responding to plaintiffs' document production requests requires BankCo to review the company's archives of electronic communications to identify all responsive email, instant messages, and other electronic documents. The accuracy of this process is very important, as BankCo is required to produce all non-privileged documents that are responsive to the requests. It must also remove any privileged documents, such as attorney-client confidential communications and attorney work product, prior to producing responsive documents to plaintiffs. BankCo would like to have an accurate and reliable method to compile an initial set of responsive documents so it can save the considerable time and expense of a completely manual review.

To handle the first category of requested documents, the Galaxy system must first identify all the relevant terms (or synonyms) corresponding to TechnologyCo, FoodCo, and PharmaCo, 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 (i.e., from January 1, 2005 to present), the sender and recipient list (the three hedge fund customers), and a list of terms regarding company TechnologyCo, FoodCo, and PharmaCo, against the body and attachment. To filter out any privileged documents, we add all attorney email addresses as negation terms to the sender and recipient list. However, this may not guarantee that we have the complete list or filter out all privileged information. Thus, manual review of the initial result set by legal personnel is required to confirm the documents to be produced.

The second category is similar to the first one, which also includes any search term corresponding to the three hedge 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 hedge fund 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 category of requested documents (i.e., brokers interfacing with HedgeA, HedgeB, or HedgeC), 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, $[\text{first}(e^{v_i-v_j}), \text{last}(e^{v_i-v_j})]$, which contains the date of first communication, $\text{first}(e^{v_i-v_j})$, and the date of the last communication, $\text{last}(e^{v_i-v_j})$, 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 $\text{last}(e^{v_l-v_k})$ is equal or later than the earliest start date, $\min(\text{first}(e^{v_l-v_{i+1}}))$, of any reachable path, $(v_1, \dots, v_i, v_{i+1}, \dots, v_j)$, 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 results, Galaxy lessens the amount of manual review needed to compile the responsive documents.

5.2 Internal Investigation Discovery

An internal investigation discovery request may be triggered, for example, 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. But 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 typically 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 material, nonpublic 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 M&A department, containing the name of clients AirlineCo, ConstructionCo, or AutoCo, 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 could 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, such as "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 Galaxy's Discovery System

Figure 3 shows the main screen of Galaxy's discovery component, which is logically divided 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 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 current and previous client history and transactions executed by the selected person around specific dates.

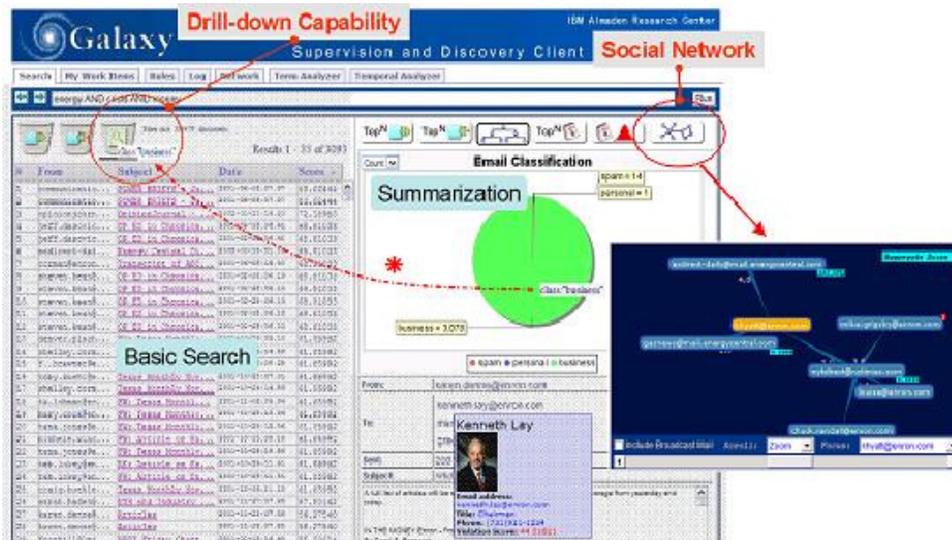


Fig. 3. Galaxy Discovery Interface

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 [12] (which provide visual representations of words based on the frequency) and stack graph, as seen in the Many Eyes project [13]. Users can further refine the search by drilling down (i.e., dragging and dropping the refined dimension into the drill-down basket, as shown in

Figure 3) 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 [14], CA Message Manager (previously known as iLumin) [15] and Zantaz [16]. 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 (e.g., evidence categories), which are easier to maintain, especially if the rule base is large (e.g., several thousand rules).

On the discovery side, commercial products, such as Zantaz [16], Symantec Enterprise Vault [17] and ZipLip [18], 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 [19][20], 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 [21] and SONAR [22] are fraud detection systems developed by NASD which use a variety of artificial intelligence 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 various electronic communications [23]. 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 strengths and weaknesses of its assumptions and approach. One potential area for future research is to develop supervision and discovery technologies that preserve individual privacy. For instance, a privacy-preserving supervision system could obscure certain private information in intercepted email communications. Similarly, improved internal 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 would be 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 material, nonpublic information. However, this activity is not detected by the supervision system because the tips are non-obvious and the resulting trades, in isolation, are small. If this communication pattern could be traced and the trades were considered in the aggregate, the financial impact would be significant. Further research is necessary to detect such patterns in electronic communication archives.

8 Conclusion

Supervision and discovery capabilities are critical for information systems used by financial services companies. In this paper, we introduced Galaxy, which provides advanced supervision and discovery of electronic communications, and discussed its advantages over other available technologies.

Galaxy empowers companies to enforce supervisory policies and procedures, conduct internal investigations, and efficiently search electronic archives to respond to legal discovery requests. 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.

Although this paper describes application scenarios from the financial services industry, Galaxy can also be applied to other industries requiring supervision or discovery capabilities. We hope this work will assist the research community in

developing more convenient and useful methods of managing electronic communications.

References

1. NASD Rule 3010, <http://finra.complinet.com/finra/>
2. NYSE Rule 342, http://rules.nyse.com/NYSE/NYSE_Rules/
3. FINRA Regulatory Notice 07-59, http://www.finra.org/web/groups/rules_regs/documents/notice_to_members/p037553.pdf
4. 15 U.S.C. Section 78o, <http://sec.gov/about/laws/sea34.pdf>
5. SEC Rule 10b-5, <http://www.sec.gov/rules/final.shtml>
6. Hearst, M. "Next Generation Web Search: Setting Our Sites," *IEEE Data Eng. Bull.* 23(3): 38-48 (2000).
7. Weiss, S.M., Indurkha, N., Zhang, T., and Damerau, F. "Text Mining: Predictive Methods for Analyzing Unstructured Information," Springer, March 2007, 236 pages, ISBN 0387954333.
8. Unstructured Information Management Architecture (UIMA), <http://incubator.apache.org/uima/>
9. 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.
10. Levenshtein, I. V. "Binary codes capable of correcting deletions, insertions, and reversals," *Doklady Akademii Nauk SSSR*, 163(4):845-848, 1965.
11. Landauer, T., Foltz, P.W., and Laham, D. "Introduction to Latent Semantic Analysis". *Discourse Processes* 25: 259-284. 1998.
12. Hassan-Montero, Y., and Herrero-Solana, V., "Improving tag-clouds as visual information retrieval interfaces," *Proc. InfoSciT 2006*.
13. Many Eyes, <http://services.alphaworks.ibm.com/manyeyes/home>
14. Orchestria, <http://www.orchestria.com/>
15. CA Message Manager, <http://www.ca.com/us/products/product.aspx?ID=5707>
16. Zantaz, <http://www.zantaz.com/>
17. Symantec Enterprise Vault, http://www.symantec.com/enterprise/products/overview.jsp?pcid=1018&pvid=322_1
18. ZipLip, <http://www.ziplip.com/>
19. 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.
20. Stolfo, S., Creamer, G., and Hershkop, S. "A Temporal Based Forensic Analysis of Electronic Communication," 2006, *Digital Government Proceedings*, San Diego, CA.
21. 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.
22. Goldberg, H., Kirkland, J., Lee, D., Shyr, P., and Thakker, D. "The NASD Securities Observation, News Analysis and Regulation System (SONAR)". *Proc. of IAAI03*.
23. Phua, C., Lee, V., Smith, K., and Gayler, R. "A comprehensive survey of data mining-based fraud detection research," *Artificial Intelligence Review*, 2005.