Visual Data Mining: Recognizing Telephone Calling Fraud

Kenneth C. Cox, Stephen G. Eick*, and Graham J. Wills
{kcc,eick,gwills}research.bell-labs.com
Bell Laboratories/Lucent Technologies

Ronald J. Brachman
AT&T Laboratories

17 April 1997

Abstract

Human pattern recognition skills are remarkable and in many situations far exceed the ability of automated mining algorithms. By building domain-specific interfaces that present information visually, we can combine human detection with machines far greater computational capacity. We illustrate our ideas by describing a suite of visual interfaces we built for telephone fraud detection.

1 Introduction

One way to improve our ability to make use of large, complex, information-rich data sets is to build on the fact that people are at the heart of the data mining enterprise. Human pattern recognition skills are remarkable, and far exceed the ability of any existing technology to detect interesting patterns and relevant anomalies. In particular, by properly taking advantage of peoples' abilities to deal with visual presentations, we may revolutionize the way we understand large amounts of data.

Researchers at Bell and AT&T Laboratories have worked to exploit the pattern detection capabilities of the human visual system by building a suite of tools and applications that flexibly encode data using color, position, size, and other visual characteristics [EF96]. Multiple different views of the same data can be interlinked, so a change in one view is reflected instantaneously in

*Bell Laboratories/Lucent Technologies, Room 1G-351, 1000 East Warrenville Road, Naperville, IL 60566

the others. Applications using this technology have been used to solve a wide variety of business-critical problems, including:

- improving programmer productivity by visualizing change patterns in large software systems [BE96];
- showing dynamic program slices and code coverage;
- manipulating the integrity constraints in huge relational databases [AEP96];
- visualizing computer log files, audit trails and other time-stamped data;
- analyzing network traffic flows [BEW95];
- understanding communities of interest in networks;
- displaying customer behavior signatures;
- segmenting customers into markets;
- understanding customer retention patterns.

Recently Lucent launched a software venture, Visual Insights^{TM}, to bring productized versions of these applications to the commercial market[1]. One of the most interesting applications of visual data mining involves fraud detection and investigation.

2 Telecommunications Fraud

Telecommunications fraud is an industry-wide problem. The percentage of fraudulent calls is small with respect to overall call volume, but the overall cost is significant; analysts estimate that industry losses to fraud in the U.S. alone amount to as much as $1 billion per year. The most significant areas of loss are in wireless and international calls, because they are the most expensive. This both makes them attractive to perpetrators of fraud (known variously as "bandits", "hackers", and "fraudsters") and means that when fraud is committed the costs mount quickly.

Fraud is also dynamic; as soon as the bandits realize they are in danger of being detected, they invent ways to circumvent security measures. Visualization techniques that rely on peoples' ability to detect anomalies and which are provided with close-to-real-time data feeds give us some hope of getting ahead of the fraud problem. The idea is that while machine-based detection methods are largely static, the human visual system is dynamic and can easily adapt to the ever-changing techniques used by the bandits.

The raw information available for calling fraud detection is a real-time feed of call detail data. For each phone call an AMA billing record, a telecommunications industry data format standard, is created and stored in huge network databases. These records contain detailed information about the call such as its time and duration, the caller and called number, which telecommunications services were used, and (for international calls) the destination country. Although the primary use of the AMA records by the phone companies is to generate bills, this data stream can also be mined to detect and investigate fraud.

[1] http://www.bell-labs.com/project/visualinsights/

The visualization approach to detecting international calling fraud involves a display of calling activity that lets the user quickly see unusual patterns. Then, using one or more drill-down views, suspicious patterns may be further investigated.

Figure 1 uses *NicheWorks* to show international calling patterns during an eight hour period to forty-one separate countries. Each country is represented by an unfilled circle, and each subscriber by a filled circle with the size and color encoding the total number of calls made by the subscriber. If a subscriber called a country during the eight hour period, the two are connected by a line, with the width and color of the line encoding the total number of calls made between the two endpoints. Most nodes and links are small since the vast majority of customers make few international calls, but the large brightly colored nodes indicate something worth further investigation.

In Figure 2 the user has interactively filtered away the small nodes and links, zoomed in on a suspicious calling pattern, and selected a subscriber node. The node is displayed in a detail view at the upper left of the figure. This view shows that during the eight hour period the subscriber made 126 calls to certain Middle East and Latin American countries. This pattern was later confirmed as fraud.

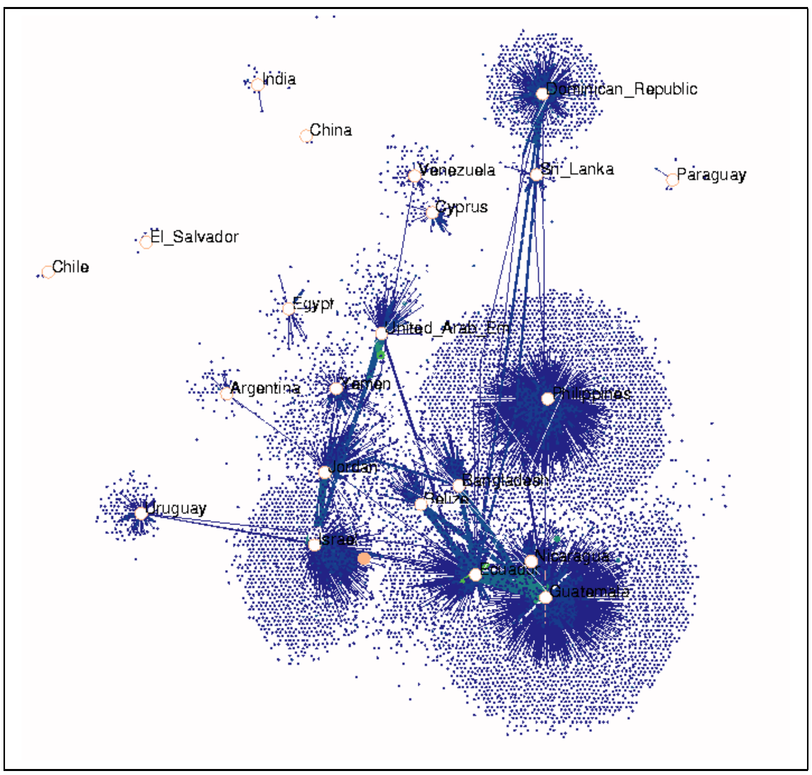

Figure 1: A NicheWorks display showing eight hours of international calling. The nodes represent the subscribers and the lines from the nodes to the countries encode the total number of calls. The large brightly colored nodes are unusual and further investigation revealed that some of them were involved in fraud.

The NicheWorks visualizations in Figures 1 and 2 use rollups of the AMA databases to provide a high-level summary of calling patterns and usage, but

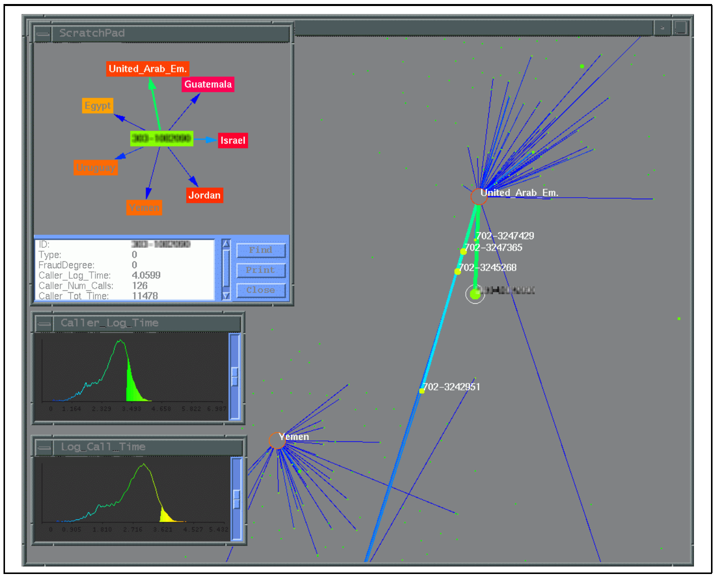

Figure 2: A suspicious calling pattern, later confirmed as fraud, for a subscriber who made 126 international calls during an eight hour period over a holiday weekend. *The telephone numbers in this figure have been modified.*

when investigating fraud more detail is often required. Figure 3 shows a *SeeCalls*[2] view, which can be obtained by selecting a node from the NicheWorks view and requesting a drill-down. The AMA records for the selected node are extracted from the database and displayed.

In this case 97 AMA records were extracted, representing calls made during a four hour period on August 18, 1995. These calls were made to destinations around the world from a single, multi-line phone system.

The main view is similar to a scatterplot, where the horizontal axis represents the time the call began and the vertical axis the duration of the call (calls exceeded the user-selected threshold of 57 minutes are shown with an arrow at the top). Many of the calls are overlapping, indicating several simultaneous users of the originating phone system, and all were made very early in the morning. The calls are color-coded by the destination country, as shown in the subsidiary barplot at the right of the figure. Many different countries were involved. The other barplots show the billing type or `BNC` (all of these calls were billed to third parties) and the type of the termination. Two of the calls were to offshore audiotex lines (often sex lines), a favorite target of bandits.

The "visual signature" shown in Figure 3 is highly characteristic of fraud, and these 97 calls were indeed subsequently determined to be fraudulent.

3 Discussion

Visual data mining complements machine-oriented data mining techniques in four ways. First, people excel at detecting patterns. Unfortunately, we tire easily, have limited memory, and have short attention spans making us unsuitable

[2]The initial S-based static version of *SeeCalls* was developed by Rick Becker, Linda Clark and Diane Lambert.

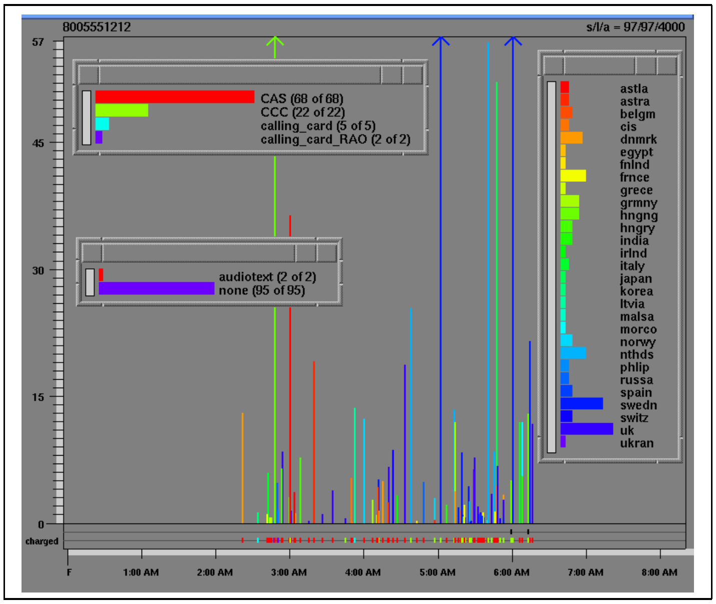

Figure 3: A detail view of AMA record data, showing 97 suspicious calls which were later confirmed as fraud.

for routine repetitive tasks. On the other hand, computers excel at such tasks, while pattern detection for computers is difficult and often requires sophisticated algorithms and delicate software. By combining the human ability to identify patterns with the computer's capability to iteratively and tirelessly search for like instances across massive datasets, we leverage the abilities of both.

Secondly, the dynamic nature of fraud makes it a particularly challenging detection problem for static algorithms. Bandits quickly adapt to fixed threshold-based detection systems; as soon as one type of fraud is detected and stopped, they attempt to discover a new one. In the visual approach we are looking for *any* unusual patterns, making it an effective strategy for identifying new classes of fraud.

Thirdly, in many ways the customer calling patterns for bandits are similar a service provider's best customers – both classes of communications systems user make lots of expensive calls! Since the primary remedy for fraud is service deactivation, which could have dire consequences if misapplied to a legitimate customer, it is important that a person be involved. Thus, building user interfaces that increase technician engagement improves the efficiency of the whole fraud prevention system.

Finally, the immediacy of the visualizations, and the complete control over dimensions of display make this technology a significant complement to other more traditional data mining tools.

There are two key innovations embodied in our visual fraud detection and investigation system. First, we provide a visual metaphor for representing calling communities, shown in Figures 1 and 2. It is well-known that calling communities form a directed graph with the nodes and links corresponding to the subscribers and calls, respectively. Our innovation was realizing that we could usefully display calling graphs with hundreds of thousands of nodes and links. We achieve this in NicheWorks using a graph layout based on a potential minimization algorithm [EW93] [BETT94].

The second key innovation involves the novel transaction-oriented display metaphor of *SeeCalls* illustrated in Figure 3. Using this display we easily show the salient characteristics of several hundred calls.

By combining the two displays and using *SeeCalls* as a drill-down view accessible through NicheWorks, we produce a very useful tool for the detection and investigation of fraud. This combination was facilitated because both were built on top of a common substrate, Bell Labs' *Vz* software library. *Vz* is an object-oriented cross-platform C++ library which encapsulates interaction, graphics, and view linking and provides the core objects and data structures for visualizations.

4 Conclusions

Visual data mining involves inventing a graphical representation for a complex data sets and building a user interface to manipulate the representation in search of patterns. In this paper we describe two representations, one for calling communities and the other for showing individual calls, that have proven effective for telecommunications fraud detection. The key idea is realizing people complement machines and by building visual user interfaces, we can better exploit the capabilities of both for knowledge discovery.

5 Acknowledgments

We wish to thank Richard Becker, Diane Lambert, and Linda Clark for their assistance and suggestions during the development of *SeeCalls* . We also wish to acknowledge the contributions of the Visual Insights^{TM} development team, for converting a research prototype into a working fraud investigation system.

References

[AEP96] Jacqueline M. Antis, Stephen G. Eick, and John D. Pyrce. Visualizing the structure of relational databases. *IEEE Software*, pages 72–79, January 1996.

[BE96] Thomas A. Ball and Stephen G. Eick. Software visualization in the large. *IEEE Computer*, 29(4):33–43, April 1996.

[BETT94] G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis. Algorithms for drawing graphs: An annotated bibliography. *Computational Geometry Theory and Applications*, 4:235–282, 1994.

[BEW95] Richard A. Becker, Stephen G. Eick, and Allan R. Wilks. Visualizing network data. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):16–28, March 1995.

[EF96] Stephen G. Eick and Daniel E. Fyock. Visualizing corporate data. *AT&T Technical Journal*, 75(1):74–86, January/February 1996.

[EW93] Stephen G. Eick and Graham J. Wills. Navigating large networks with hierarchies. In *Visualization '93 Conference Proceedings*, pages 204–210, San Jose, California, 25-29 October 1993.