3

Data Mining

Revealing its origins and widespread use in business, data mining goes by many names, including knowledge management, knowledge discovery, and sense making.[1] Data mining is "[a]n information extraction activity whose goal is to discover hidden facts contained in databases."[2] In other words, data mining involves the systematic analysis of large data sets using automated methods. By probing data in this manner, it is possible to prove or disprove existing hypotheses or ideas regarding data or information, while discovering new or previously unknown information. In particular, unique or valuable relationships between and within the data can be identified and used proactively to categorize or anticipate additional data. Through the use of exploratory graphics in combination with advanced statistics, machine learning tools, and artificial intelligence, critical "nuggets" of information can be mined from large repositories of data.

Is Data Mining Evil?

Further confounding the question of whether to acquire data mining technology is the heated debate regarding not only its value in the public safety community but also whether data mining reflects an ethical, or even legal, approach to the analysis of crime and intelligence data. The discipline of data mining came under fire in the Data Mining Moratorium Act of 2003.

Unfortunately, much of the debate that followed has been based on misinformation and a lack of knowledge regarding these very important tools. Like many of the devices used in public safety, data mining and predictive analytics can confer great benefit and enhanced public safety through their judicious deployment and use. Similarly, these same assets also can be misused or employed for unethical or illegal purposes.

One of the harshest criticisms has addressed important privacy issues. It has been suggested that data mining tools threaten to invade the privacy of unknowing citizens and unfairly target them for invasive

investigative procedures that are associated with a high risk of false allegations and unethical labeling of certain groups. The concern regarding an individual's right to privacy versus the need to enhance public safety represents a long-standing tension within the law enforcement and intelligence communities that is not unique to data mining. In fact, this concern is misplaced in many ways because data mining in and of itself has a limited ability, if any, to compromise privacy. Privacy is maintained through restricting access to data and information. Data mining and predictive analytics merely analyze the data that is made available; they may be extremely powerful tools, but they are tools nonetheless. With data mining, ensuring privacy should be no different than with any other technique or analytical approach.

Unfortunately, many of these fears were based on a misunderstanding of the Total Information Awareness system (TIA, later changed to the Terrorism Information Awareness system), which promised to combine and integrate wide-ranging data and information systems from both the public and private sectors in an effort to identify possible terrorists. Originally developed by the Defense Advanced Research Projects Agency (DARPA), this program was ultimately dismantled, due at least in part to the public outcry and concern regarding potential abuses of private information. Subsequent review of the program, however, determined that its main shortcoming was related the failure to conduct a privacy impact study in an effort to ensure the maintenance of individual privacy; this is something that organizations considering these approaches should include in their deployment strategies and use of data-mining tools.

On the other hand, some have suggested that incorporation of data mining and predictive analytics might result in a waste of resources. This underscores a lack of information regarding these analytical tools. Blindly deploying resources based on gut feelings, public pressure, historical precedent, or some other vague notion of crime prevention represents a true waste of resources. One of the greatest potential strengths of data mining is that it gives public safety organizations the ability to allocate increasingly scarce law enforcement and intelligence resources in a more efficient manner while accommodating a concomitant explosion in the available information—the so-called "volume challenge" that has been cited repeatedly during investigations into law enforcement and intelligence failures associated with 9/11. Data mining and predictive analytics give law enforcement and intelligence professionals the ability to put more evidence-based input into operational decisions and the deployment of scarce resources, thereby limiting the potential waste of resources in a way not available previously.

Regarding the suggestion that data mining has been associated with false leads and law enforcement mistakes, it is important to note that these errors happen already, without data mining. This is why there are so many checks and balances in the system—to protect the innocent. We do not need data mining or technology to make errors; we have been able to do that without the assistance of technology for many years. There is no reason to believe that these same checks and balances would not continue to protect the innocent were data mining to be used extensively. On the other hand, basing our activities on real evidence can only increase the likelihood that we will correctly identify the bad guys while helping to protect the innocent by casting a more targeted net. Like the difference between a shotgun and a laser-sited 9mm, there is always the possibility of an error, but there is much less collateral damage with the more accurate weapon.

Again, the real issue in the debate comes back to privacy concerns. People do not like law enforcement knowing their business, which is a very reasonable concern, particularly when viewed in light of past abuses. Unfortunately, this attitude confuses process with input issues and places the blame on the tool rather than on the data resources tapped. Data mining can only be used on the data that are made available to it. Data mining is not a vast repository designed to maintain extensive files containing both public and private records on each and every American, as has been suggested by some. It is an analytical tool. If people are concerned about privacy issues, then they should focus on the availability of and access to sensitive data resources, not the analytical tools. Banning an analytical tool because of fear that it will be misused is similar to banning pocket calculators because some people use them to cheat on their taxes.

As with any powerful weapon used in the war on terrorism, the war on drugs, or the war on crime, safety starts with informed public safety consumers and well-trained personnel. As is emphasized throughout this text, domain expertise frequently is the most important component of a well-informed, professional program of data mining and predictive analytics. As such, it should be seen as an essential responsibility of each agency to ensure active participation on the part of those in the know; those professionals from within each organization that know where the data came from and how it will be used. To relinquish the responsibility for analysis to outside organizations or consultants should be viewed in the same way as a suggestion to entirely contract patrol services to a private security corporation: an unacceptable abdication of an essential responsibility.

Unfortunately, serious misinformation regarding this very important tool might limit or somehow curtail its future use when we most need it in our fight against terrorism. As such, it is incumbent upon each organization to ensure absolute integrity and an informed decision-making process regarding the use of these tools and their output in an effort to ensure their ongoing availability and access for public safety applications.

3.1 Discovery and Prediction

When examining drug-related homicide data several years ago, we decided to experiment with different approaches to the analysis and depiction of the information. By drilling down into the data and deploying the information in a mapping environment, we found that the victims of drug-related homicides generally did not cross town to get killed. While it makes sense in retrospect, this was a very surprising finding at the time. This type of analysis of homicide data had not been considered previously, although after it had been completed it seemed like a logical way to view the information.

After further analysis of the data, we were able to generate a prediction regarding the likely location and victim characteristics of one of the next

incidents. Within the next twelve hours, a murder was committed with characteristics that were strikingly similar to those included in the prediction, even down to the fact that the victim had not crossed town to get killed.

This embodies the use of data mining and predictive analytics in law enforcement and intelligence analysis. First, the behavior was characterized and, through this process, new information was "discovered." The idea of looking at the information in this fashion to determine the relationship between the victim's residence and subsequent murder location made sense, but had not been done before. Adding value to crime information in this manner deviates significantly from the traditional emphasis on counting crime and creating summary reports. By looking at the data in a different way, we were able to discover new facets of information that had significant operational value.

Second, by characterizing the behavior, it could be modeled and used to anticipate or predict the nature of future events. The ability to anticipate or predict events brings a whole new range of operational opportunities to law enforcement personnel. Much as in the movie *Minority Report*, once we can anticipate or predict crime, we will have the ability to prevent it. Unlike the movie, however, crime prevention can be effected through the use of proactive deployment strategies or other operational initiatives, rather than proactive incarceration of potential offenders. On the other hand, the ability to characterize risk in potential victims provides an opportunity for targeted, risk-based interventions that ultimately can save lives and provide safer neighborhoods for all, a topic that will be covered in Chapter 11.

This example, although a somewhat odd and inelegant use of "brute force" analytics, embodies the essence of data mining and predictive analytics within the public safety arena. Through the use of these powerful tools, we can understand crime and criminal behavior in a way that facilitates the generation of actionable models that can be deployed directly into the operational environment.

3.2 Confirmation and Discovery

At a very simple level, data mining can be divided into confirmation and discovery. Criminal investigation training is similar to case-based reasoning.[3] In case-based reasoning, each new case or incident is compared to previous knowledge in an effort to increase understanding or add informational value to the new incident. In addition, each new incident is added to this internal knowledge base. Before long, an investigator has developed an internal set of rules and norms based on accumulated experience. These rules and norms are

then used, modified, and refined during the investigation of subsequent cases. Analysts and investigators will look for similarities and known patterns to identify possible motives and likely suspect characteristics when confronted with a new case. This information is then used to understand the new case and investigate it.

These internal rule sets also allow an investigator to select suspects, guide interviews and interrogations, and ultimately solve a case. These existing rule sets can be evaluated, quantified, or "confirmed" using data mining. In addition, internal rule sets can be modified and enhanced as additional information is added and integrated into the models. Finally, as predictive algorithms are developed, we can extend beyond the use of data mining for simple characterization of crime and begin to anticipate, predict, and even prevent crime in some cases.

Many seasoned homicide investigators can identify a motive as the call comes in, based on the nature of the call, geographic and social characteristics of the incident location, and preliminary information pertaining to the victim and injury patterns. For example, a young male killed in a drive-by shooting in an area known for open-air drug markets is probably the victim of a drug-related homicide. Additional information indicating that the victim was known to be involved in drug selling will further define the motive and suggest that likely suspects will include others involved in illegal drug markets. Post-mortem information indicating that the victim had used drugs recently before his death will add additional value to our understanding of the incident.

Law enforcement personnel, particularly those who have acquired both experience and success working on the streets, have internal "rule sets" regarding crime and criminal behavior that can be invaluable to the data mining process. In some initial research on juveniles involved in illegal drug markets, we found results that differed significantly from the prevailing opinions in the literature, which indicated that most drug sellers are involved in illegal drug markets in order to support their personal use.[4] The common scenario involved a poor individual who experimented with drugs, rapidly became hooked, escalated to "hard" drugs, and then needed to rely on drug sales to support a rapidly growing, expensive habit. Our results indicated a very different scenario. The data that we reviewed indicated that drug sellers actually functioned very well, tended to have excellent social skills, and rarely used illegal drugs beyond some recreational use of marijuana. It was only when we looked at the relatively small group of drug traffickers who had been shot previously that we found relatively high levels of substance use and generally poor functioning. Our findings were somewhat confusing until we had the opportunity to discuss them with law enforcement professionals who were still actively working illegal narcotics. These individuals

were not surprised by our findings. They pointed out to us that most successful drug dealers do not use what they sell because it cuts into their profits and, perhaps more importantly, impairs their ability to function in an extremely predatory criminal environment. Moreover, those drug dealers that do not function well generally do not live very long. From this point on, we made a point of using this type of reality testing not only to evaluate or confirm our findings but also to guide our research in this area. In many ways, this approach embodies data mining as a confirmation tool. By learning more about the internal rule sets that detectives used to investigate cases, we were able to structure and guide our data mining. In most cases we were able to confirm their instincts; however, in other cases the results were truly surprising.

3.3 Surprise

By using automated search and characterization techniques, it also is possible to discover new or surprising relationships within data. The ability to characterize large databases far exceeds the capacity of a single analyst or even a team of analysts.

The Commonwealth of Virginia has been a pioneer in the use of DNA databases to identify suspects and link cases based on the use of DNA evidence. Having achieved considerable success in this area, the Commonwealth boasts a record of approximately one "cold hit" per day.[5] One noteworthy feature of the Virginia database is that it includes DNA from all convicted felons, as opposed to only those known to be violent or sexually violent. An informal conversation with the director of forensic sciences revealed that a large number of their DNA cold hits had come from offenders with no prior history of either violent or sex-related offenses. Many of these offenders had been incarcerated previously for property crimes, particularly burglary. This was a particularly surprising finding because it had been assumed that most of the cold hits would come from offenders previously convicted of sexual or violent crimes. In fact, some states had restricted their inclusion criteria to only those felons convicted of violent or sexually related crimes. The assumption was that these would be the only individuals of interest because they would be the most likely to recidivate in a violent manner.

Having spent considerable time reviewing the case materials associated with murderers, we could recall anecdotally several cases where a specific offender escalated from nonviolent to violent offending. Perhaps most noteworthy was the Southside Strangler case in Virginia, which subsequently became the first case to use DNA evidence to convict a suspect in court. Prior to committing

several horrific murders in Richmond, Virginia, Timothy Spencer was known to have committed burglaries in northern Virginia.

Challenged by this seemingly spurious finding, we embarked on an analysis of several large correctional databases to determine whether there was something unusual about the sample of DNA cold hits that could explain this apparent anomaly, or whether it was real. Using discriminant analysis, a classification technique, it was determined that a prior burglary was a better predictor than a prior sex offense of a subsequent stranger rape, a very surprising finding. Subsequent review of the sex offender literature confirmed our findings.

It is important to note, however, that in many cases the type of nonviolent offending was different than crimes perpetrated by offenders who did not escalate. The use of data mining to identify and characterize "normal" criminal behavior has turned out to be an extremely valuable concept and is discussed in detail in Chapter 10.

3.4 Characterization

Using data mining, we can begin to further characterize crime trends and patterns, which can be essential in the development of specific, targeted approaches to crime reduction. For example, we know that violence can take many forms, which are addressed through different approaches. This is the first step in the modeling process. A program to address domestic violence might employ social service workers as second responders to incidents of domestic violence. Victim education, offender counseling, and protective orders also might be implemented. Drug-related violence, on the other hand, requires a different approach. In fact, different types of drug-related violence will require different solutions, depending on their specific nature. By delving into the data and identifying associated clusters or groups of crime, we can gain additional insight into the likely causes. Ultimately, this facilitates the identification and development of meaningful, targeted intervention strategies.

Analysis of the data in this manner does not involve the use of a crystal ball. Rather, it requires an understanding of the data and the domain expertise necessary to know when, where, and how to dig into the data, what data to use, and what questions to ask about it. The importance of solid domain expertise cannot be overstated. Without knowing what has value and meaning to an understanding of the data within the context of crime and intelligence analysis, processing the information and investigating it will add little meaning and might result in bogus findings.

3.5 "Volume Challenge"[6]

Although this phrase first emerged during the period immediately after the events of September 11, the law enforcement and intelligence community have been trying to address staggering increases in data and information for many years. The number of tips, reports, complaints, and other public safety–related information confronting law enforcement and intelligence professionals on a daily basis is phenomenal. This particular information challenge can be illustrated well by following major case investigations that have been in the news.

On January 9, 2003, KXTV reported that investigators had received more than 2600 tips in response to the Laci Peterson disappearance.[7] Considering that Ms. Peterson was reported missing on Christmas Eve, the local authorities received these 2600 tips in less than 17 days, or approximately 162 tips per day, assuming that the rate of tips was distributed uniformly, which is unlikely. Similarly, during the D.C. sniper investigations, tips were being received at rates as high as 1000 per hour.[8] Given the nature of these crimes and the volumes of associated information, perhaps the most important task associated with crime tips is logging them into a database in some sort of systematic fashion. This challenge is followed closely by the need to analyze or make some sense out of the information, identifying and clustering those that are similar, and at the same time highlighting any patterns and trends in the information.

How do we even begin to analyze this volume of information, though? In many cases, the tips are initially logged and then shared as leads with investigators. In some cases, tip information might be maintained in electronic databases but, even under the best of circumstances, automated search and analysis is limited by available analytical tools and capacity. Until recently, it was almost impossible for a single analyst or even an analytical task force to thoroughly review and assimilate this amount of information in any sort of meaningful or systematic fashion. Unfortunately, this approach significantly limits the value of tips, which ultimately can compromise public safety and cost lives.

For example, subsequent review of the D.C. sniper investigation revealed that the actual vehicle used by the snipers had been seen and reported. In other words, many of the answers to solving the case resided in the tip databases, but the volume of information precluded their detection. The key nuggets of information essential to identifying a suspect or impending event often are identified in retrospect, which frequently is too late. As the review of high-profile cases often reveals, the information necessary to closing a case or preventing a

tragedy might be hidden in plain sight within the large, unmined tip databases residing in law enforcement and intelligence organizations throughout this country and throughout the world. Unfortunately, as time passes and the number of tips continues to grow, the ability to efficiently and effectively review and analyze the information using traditional methodologies decreases concomitantly. That is why it is so important that public safety agencies adopt and employ the automated search strategies and data mining techniques that are now available.

Clearly, no case will be decided exclusively based on the use of computer programs and analytics, but these tools can be brutally objective, beyond the most seasoned detective. Data mining can also transcend the media reports, focus, hype, information overload, and even the brutality and violence associated with the crime scene, focusing exclusively on the compiled information and facts. As such, it offers a tremendous advantage to the public safety community over traditional methodologies.

3.6 Exploratory Graphics and Data Exploration

Available Software

As has been said more than a few times throughout this text, the need to increase the analytical capacity in the crime and intelligence community within the United States, coupled with increasing interest in the area of data mining, has supported a flurry of new products and even some renamed old products. Therefore, one goal of this text is to create an informed consumer, for two reasons.[9] First, and perhaps most obvious, is that data mining software can be very expensive. It is important, therefore, to ensure that you are getting what you have paid for.

Second, and perhaps more important, is the fact that most readers of this text will be considering a purchase in support of some sort of public safety application. Whether for crime or intelligence analysis, it is extremely important to ensure that the outcomes are reliable and valid. An inferior product or one that does not have the analytical muscle to back its advertisements represents a failure not only in purchasing and budgetary decisions but also in the support of public safety. In other words, an error in this realm can cost not only very scarce dollars, it can also cost lives.

It is unlikely that every agency will need to purchase the most expensive and high-powered tools available. Rather, some consideration should be given to the

nature of the need within the organization and the best analytical approach and associated tools for the job. One good place to start might be the development of information-based deployment strategies because, if used appropriately, data mining tools can pay for themselves relatively quickly in personnel savings alone. Another area to consider is investing some time and effort into information management, which can facilitate the full exploitation of predictive analytics. Ideally, systems designed specifically to deploy this information through mapping or directly into the analytical environment will continue to be developed and enhanced. It is not necessary to create an analytical unit that is outfitted to look like mission control. Identifying some initial, manageable areas for improvement that can be enhanced and expanded over time represents best practice in the acquisition of new technologies.

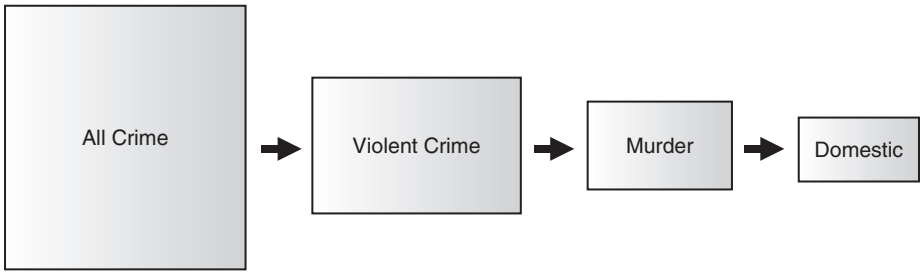
One particularly interesting market forecast indicates that an area of growth in the data mining field includes specialty niche markets.[10] Products developed for these niche markets will be tailored toward domain experts. These analytical tools will require less technical expertise and training, relying instead on the end user's knowledge of their field and the use of innovative graphical interfaces and other visualization techniques. The availability of tools developed specifically for law enforcement, security, and intelligence analysts promises to increase even further the availability of data mining and predictive analytics in the applied setting.

The first question might be: Do I need power tools for this? The answer is a most definite "maybe." Exploring the data to identify unique patterns and trends almost certainly requires the use of computerized approaches. The uncertainty generally involves the specific nature of the tools required. Because this is an important consideration that can impact your success with these tools, it has been addressed throughout this section in an effort to support the "informed consumer" in the acquisition process.

Figure 3-1 illustrates how it is possible to narrow the focus on the data in an effort to identify relatively homogenous subsets of information. For example, the category of "crime" is large and relatively heterogeneous. Included within this category is everything from misdemeanor theft to murder. Trying to do anything with such a diverse array of behavior is almost certain to fail; people often realize this when they try to evaluate a "crime" prevention strategy and find out later that the problem is too large for any single program to make a meaningful impact.

If we divide the data somewhat, the "violent crime" data can be selected. This is still a relatively large, heterogeneous category that is likely to include aggravated and sexual assaults as well as robberies and murders. It would be

Figure 3-1 *Narrowing the focus on the data frequently can reveal smaller groups that are relatively similar in their attributes, which facilitates subsequent analysis.*



difficult to generate many useful models or predictions on such a wide range of information. A still more detailed focus on the data provides an opportunity to add further value to a thoughtful characterization and analysis of the data.

Through further investigation of the violent crime data, another subcategory of "murder" can be identified. Again, however, this is still relatively generic. For example, a robbery-related homicide is likely to differ in many significant ways from a domestic homicide. A similar discontinuity could be revealed when domestic homicides are compared directly to drug-related homicides. Dividing "murder" based on motive or victim-perpetrator relationship will further increase the relative homogeneity of the grouping.

Why is this important? In order to accurately characterize data, so as to reveal important associations and create accurate and reliable models, it is important to generate samples that are relatively homogenous, except with regard to those unique features that have value from a modeling perspective. For example, when reviewing victims of prior firearms injuries, a first pass through the data revealed no association between the risk of being shot and the likelihood that the victim carried a weapon.[11] Dividing the sample based on the pattern of criminal offending, however, revealed an entirely different story. Those victims previously involved in aggressive or violent patterns of offending were much more likely to have been shot if they also were known to carry a weapon, which might be related to or indicative of particularly aggressive interactional patterns of behavior.

Injured drug dealers, on the other hand, were much less likely to carry a weapon, possibly indicating poor defensive skills in a very predatory criminal activity. Therefore, the association between getting shot and carrying a weapon was obscured when the data were in aggregate form. Although it initially appeared that the data were relatively homogeneous in that they were

confined to juvenile offenders, important relationships within the data were not revealed until it had been analyzed further. In this case, the associated pattern of offending was an important factor in determining the relationship between sustaining a firearms-related injury and weapon possession.

Officer Safety

Characterization of different victim risk patterns also has officer safety implications. Anecdotal reports link weapon selection to the reason for using a weapon. For example, those criminals electing to carry a weapon as an extension of an aggressive or violent approach to the world are more likely to select a weapon that is similarly menacing. These offenders frequently prefer something large and scary-looking with an increased capacity. They are willing to compromise accuracy in an effort to acquire something that fits their perceived image and lifestyle.

On the other hand, criminals electing to carry a weapon for defensive purposes, such as those involved in illegal drug markets, generally prefer something that can be readily concealed and is reliable. After being shot, a 15-year-old drug dealer revealed his decision-making process for weapon selection. Interestingly, many of the factors that he considered were similar to the ones cited by a large federal agency that had recently switched manufacturers and chosen the same brand that this juvenile drug seller had selected.

Aside from being ironic, this finding has significant implications for officer safety. By characterizing the likely weapon selection process, operational personnel are able to gain added insight into what they might encounter on the street when confronting a particular type of offender. The knowledge that many young, violent offenders are willing to select form over function while juvenile drug sellers have a preference for easily concealed, very reliable weapons has the potential to determine which party is likely to walk away from a violent encounter. To quote Miguel de Cervantes, "forewarned is forearmed, to be prepared is half of the victory."

Why do we care about this? First, these findings have direct implications for treatment programs. Given that the differences noted were behavioral patterns and styles associated with the risk for injury, it would be inappropriate to develop a generic "firearms injury survivor" group. Just as it would be crazy to create a "drug-involved offenders" group that included drug dealers as well as drug users, it would be similarly risky to combine victims who likely had been shot because of an extremely aggressive interactional style with those who had been shot as a result of poor defensive skills. These findings also have officer safety implications, in that foreknowledge of an offender's likely weapon preference can be extremely valuable on the street.

3.7 Link Analysis[12]

Sometimes we want to ask the question, "What things go together?" Typically, these might be features of a crime such as where and when the crime occurred, what types of property, people, or vehicles were involved, the methods used, and so on. One way of answering such "link analysis" questions is to use web graphs, which show associations between items (such as individuals, places, or any other element of interest) by points in a diagram, with lines depicting the links between them. These tools can have added value in that the strength of the association can be depicted by the strength of the line. For example, a solid line conveys a much stronger relationship than a dashed line.

One common pitfall in link analysis is to over-interpret the identified relationships or results, particularly those with unequal distributions. This issue is illustrated further in Chapter 1, but the best way around this issue, like most others in data mining, is to know your domain and know your data. It always is extremely important to explore the data initially and interpret any results cautiously. Potential options for addressing this can include the use of percentages rather than the actual frequencies. Again, this is illustrated in greater detail in Chapter 1.

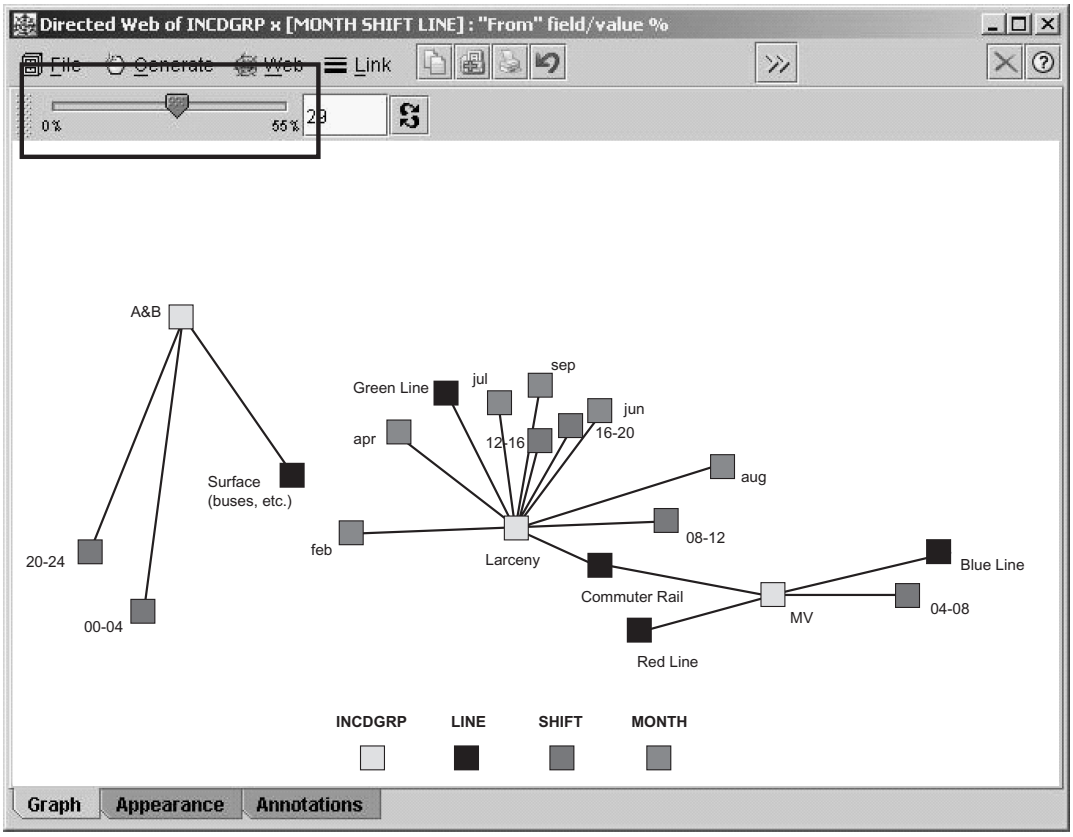
Many software tools provide a toggle option or sliding scale that gives the analyst the opportunity to adjust the thresholds used to determine the relative strengths of multiple relationships (Figure 3-2). This can be a tremendous tool, particularly during the exploration process, as it allows the analyst to reduce the "noise" so that the important relationships can be visualized easily. Other products allow the user to adjust whether the strength of the relationship is based on frequencies or percentages. Again, this can be a tremendous asset when evaluating and comparing relationships in the data.

3.8 Nonobvious Relationship Analysis (NORA)[13]

Unfortunately, life generally is not as simple as a web graph or link analysis would indicate. For example, it is not unusual for a suspect to intentionally alter the spelling of his name or attempt to vary her identity slightly in an effort to avoid detection. Similarly, Richard, Dick, Rick, Rich, Ricky, etc., all are legitimate variations of the same name. This challenge becomes even more of an issue in investigative and intelligence databases where the information can be even less uniform and reliable.

In addition, it is not uncommon in crime, particularly in organized crime or terrorism, for individuals to try to avoid having a direct relationship with other

Figure 3-2 *This figure depicts a web graph. The box highlights the tool used to fluidly adjust the thresholds in an effort to reveal or mask certain differences. (B. Haffey, SPSS, Inc.; used with permission.)*



members of the group or organization. In fact, the Al Qaeda handbook, which is available on the Internet, specifically advises that operatives significantly limit or avoid contact with others in the cell in an effort to reduce detection and maintain operational security. Clearly, this creates a significant limitation for the use of standard automated association detection techniques.

Recently, however, automated techniques referred to as Nonobvious Relationship Analysis, or NORA, have emerged from the gambling industry in Las Vegas. Used to identify cheaters, these tools have obvious implications for law enforcement and intelligence analysis. While not available as "off-the-shelf" software products at the time of this writing, they can identify links and relationships not readily identifiable using traditional link analysis software.

These tools also can identify subtle changes in numeric information, such as social security numbers. In many cases, these transpositions are unintentional keystroke errors. In others, however, numeric information is changed slightly to reduce the likelihood that information will be linked directly, which can be indicative of identify theft or similar types of fraud.

3.9 Text Mining

Most information that has value in law enforcement and intelligence analysis resides in unstructured or narrative format. Frequently, it is the narrative portion of a police report that contains the most valuable information pertaining to motive and modus operandi, or MO. It is in this section of the report that the incident or crime is described in the behavioral terms that will be used to link it to others in a series, to similar crimes in the past, or to known or suspected offenders. In addition, crime tip information and intelligence reports almost always arrive in unstructured, narrative format. Because it is unlikely that an informant will comply with a structured interview form or questionnaire, the onus is upon the analyst either to transcribe and recode the information or to identify some automated way to analyze it.

Until recently, this information was largely unavailable unless recoded. This is time consuming, and can alter the data significantly. Recoding generally involves the use of arbitrary distinctions to sort the data and information into discrete categories that can be analyzed. Unfortunately, many aspects of the data, information, and context can be compromised through this process.

Recent advances in text mining tools that employ natural language processing now provide access to this unformatted text information. Rather than crude keyword searches, the information pulled out through the use of text mining incorporates syntax and context. As a result, more complex concepts can be mined, such as "jumped the counter" or "passed a note," valuable MO characteristics that could be associated with takeover robberies or bank robberies, respectively.

Like the suspect debriefing scenario outlined in the Introduction, these tools promise to advance the analytical process in ways not considered until very recently. For example, the information obtained through the interview process can be inputted directly into the analysis and integrated with other narrative and categorical data. Moreover, tip databases can be reviewed, characterized, and culled for common elements, themes, and patterns. These tools promise to significantly enhance statement analysis, as they can identify common themes and patterns, such as those associated with deception or false allegation. It will

be truly exciting to see where these tools take the field of crime and intelligence analysis in the future.

3.10 Future Trends

Many of the new data mining programs are very intuitive and also incredibly fast. This will facilitate their use in both the planning and the operational environments. Policy decisions now can become information-based through the use of these tools. This capability was tested at the 2003 International Association of Chiefs of Police annual conference, where data mining tools were used during a workshop on police pursuits. Initial findings were augmented and enhanced on scene by analysis in this live environment. Questions regarding the data were answered as quickly as they were raised through the use of these extremely quick and intuitive tools. The use of these analytical approaches in a live environment promises to put more science and less fiction into law enforcement policy and planning. In addition, the potential value of these tools in real-time, operational planning is tremendous. The ability to model possible scenarios and outcomes will not only enhance operational strategy and deployment but also can save lives as potential risks are identified and characterized.

3.11 Bibliography

1. Helberg, C. (2002). Data mining with confidence, 2nd ed. SPSS, Inc., Chicago, IL.
2. Definition from Two Crows (www.twocrows.com), which is an excellent source of accurate yet easy to understand information on data mining and predictive analytics.
3. Casey, E. (2002). Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and Internet crime investigation. *Knowledge Solutions.* www.corpus-delicti.com/case_based.html
4. McLaughlin, C.R., Reiner, S.M., Smith, B.W., Waite, D.E., Reams, P.N., Joost, T.F., and Gervin, A.S. (1996). Firearm injuries among Virginia juvenile drug traffickers, 1992 through 1994 (Letter). *American Journal of Public Health*, **86**, 751–752; McLaughlin, C.R., Smith, B.W., Reiner, S.M., Waite, D.E., and Glover, A.W. (1996). Juvenile drug traffickers: Characterization and substance use patterns. *Free Inquiry in Creative Sociology*, **24**, 3–10; McLaughlin, C.R., Reiner, S.M., Smith, B.W., Waite, D.E.,

Reams, P.N., Joost, T.F., and Gervin, A.S. (1996). Factors associated with a history of firearm injuries in juvenile drug traffickers and violent juvenile offenders. *Free Inquiry in Creative Sociology, Special Issue: Gangs, Drugs and Violence*, **24**, 157–165.

5. McCue, C., Smith, G.L., Diehl, R.L., Dabbs, D.F., McDonough, J.J., and Ferrara, P.B. (2001). Why DNA databases should include all felons. *Police Chief*, **68**, 94–100.
6. Tabussum, Z. (2003). CIA turns to data mining; www.parallaxresearch.com/news/2001/0309/cia_turns_to.html
7. www.KXTV10.com (2003). Despite avalanche of tips, police stymied in Laci Peterson case. January 9.
8. Eastham, T. (2002). Washington sniper kills 8, truck sketch released. October 12. www.sunherald.com
9. For a current review and comparison of specific data mining products, go to Elder Research, Inc. (www.dataminginglab.com).
10. METAspectrum^{SM} Market Summary (2004). Data mining tools: METAspectrum^{SM} evaluation. META Group, Inc.
11. McLaughlin, C.R., Daniel, J., Reiner, S.M., Waite, D.E., Reams, P.N., Joost, T.F., Anderson, J.L., and Gervin, A.S. (2000). Factors associated with assault-related firearms injuries in male adolescents. *Journal of Adolescent Health*, **27**, 195–201.
12. Helberg, C. (2002).
13. Franklin, D. (2002). Data miners: New software instantly connects key bits of data that once eluded teams of researchers. *Time,* December 23.