COMPUTING SCIENCE

Connecting the Dots

Brian Hayes

IN THE FIVE YEARS since that wrenching Tuesday morning when hijacked aircraft sliced into the World Trade Center and the Pentagon, Americans have been living with a new undercurrent of worry and mistrust. Naturally, there's fear of further attacks. But there's also concern that measures taken to forestall such attacks could erode traditional rights and liberties. In recent months, controversy has erupted over reports that government agencies are monitoring Internet and telephone communications as well as financial transactions. Some of the surveillance programs are said to be sifting through gigantic data sets, scanning for patterns that might reveal criminal intent or activity.

The debate over these programs has focused mainly on legal and political questions. Are constitutional and statutory safeguards being respected? What about laws that bar intelligence agencies from spying on American citizens? Do the programs strike an appropriate balance between the right to privacy and the need for security? These are important issues, but I shall leave them to others. Here I want to ask a different kind of question: What can one expect to learn through such wholesale screening and data-mining operations? Do the communications patterns of terrorists have a signature so distinctive that computer algorithms can detect signs of a conspiracy amid trillions of other telephone calls or e-mail messages?

In addressing these questions I face an obvious impediment: Very little reliable information on the nature and scope of the surveillance programs has been made public. However, mathematicians and computer scientists have tackled problems very similar to those confronting an intelligence analyst trying to make sense of surveillance data. And social scientists have long taken an interest in the networks that bind people together—including networks of criminals and terrorists. Perhaps by combining insights from these fields we can make some plausible guesses.

Can the tools of graph theory and social-network studies unravel the next big plot?

The 411 on Telephone Snooping

The newly revealed surveillance programs seem to include several distinct activities. Some involve eavesdropping—listening in on telephone conversations or recording the content of Internet messages. A follow-the-money program gathers information from a banking clearinghouse. But the reports I find most intriguing mention efforts to analyze a database of telephone calls with the aim of tracing links among conspirators. The database includes no sound recordings or any other hints about what might have been said in a conversation; it merely lists the telephone numbers at the two ends of each call and gives the date and time when a call began and ended.

This "call detail" database sounded very familiar. Several years ago I had read of experiments done with a similar database—almost surely an earlier version of the one that is now said to be under government scrutiny. The experiments were tests of algorithms in the mathematical field known as graph theory, which studies network-like structures. The phone-call database was a useful test bed because it can be viewed as an enormous mathematical graph. I wrote about this work in an earlier column in *American Scientist* (January–February 2000).

Vague allusions to the database, or "call graph," appeared in the first public accounts of the new surveillance programs. Writing in *The New York Times* last December, Eric Lichtblau and James Risen noted, "[National Security Agency] technicians, besides actually eavesdropping on specific conversations, have combed through large volumes of phone and Internet traffic in search of patterns that might point to terrorism suspects." The nature of the operation became clearer in May when Leslie Cauley wrote in *USA Today* that at least three telephone companies are voluntarily supplying call-detail records to the NSA. Two of those companies later denied that they participate in the program, and *USA Today* retracted that part of the story. The third company, AT&T, has declined to comment on the substance of the report, and so has the NSA. When AT&T was sued for allegedly violating privacy statutes, the Bush administration moved to suppress the suits on the grounds that litigating the matter would reveal state secrets. As this issue of *American Scientist* goes to press, the facts remain murky.

Who Calls Whom

The NSA is the U.S. espionage service with responsibility for cryptography and "signals intelligence." Although its budget and staffing are secret, it is often said to be the largest of the U.S. intelligence agencies and also, incidentally, the largest employer of mathematicians in the United States and perhaps in the world. And it is assumed to possess prodigious computing resources.

Exploration of the call graph belongs to the branch of signals intelligence

Brian Hayes is Senior Writer for American Scientist. *Additional material related to the "Computing Science" column appears in Hayes's weblog at http://bit-player.org. Address: 211 Dacian Avenue, Durham, NC 27701. Internet: bhayes@amsci.org*

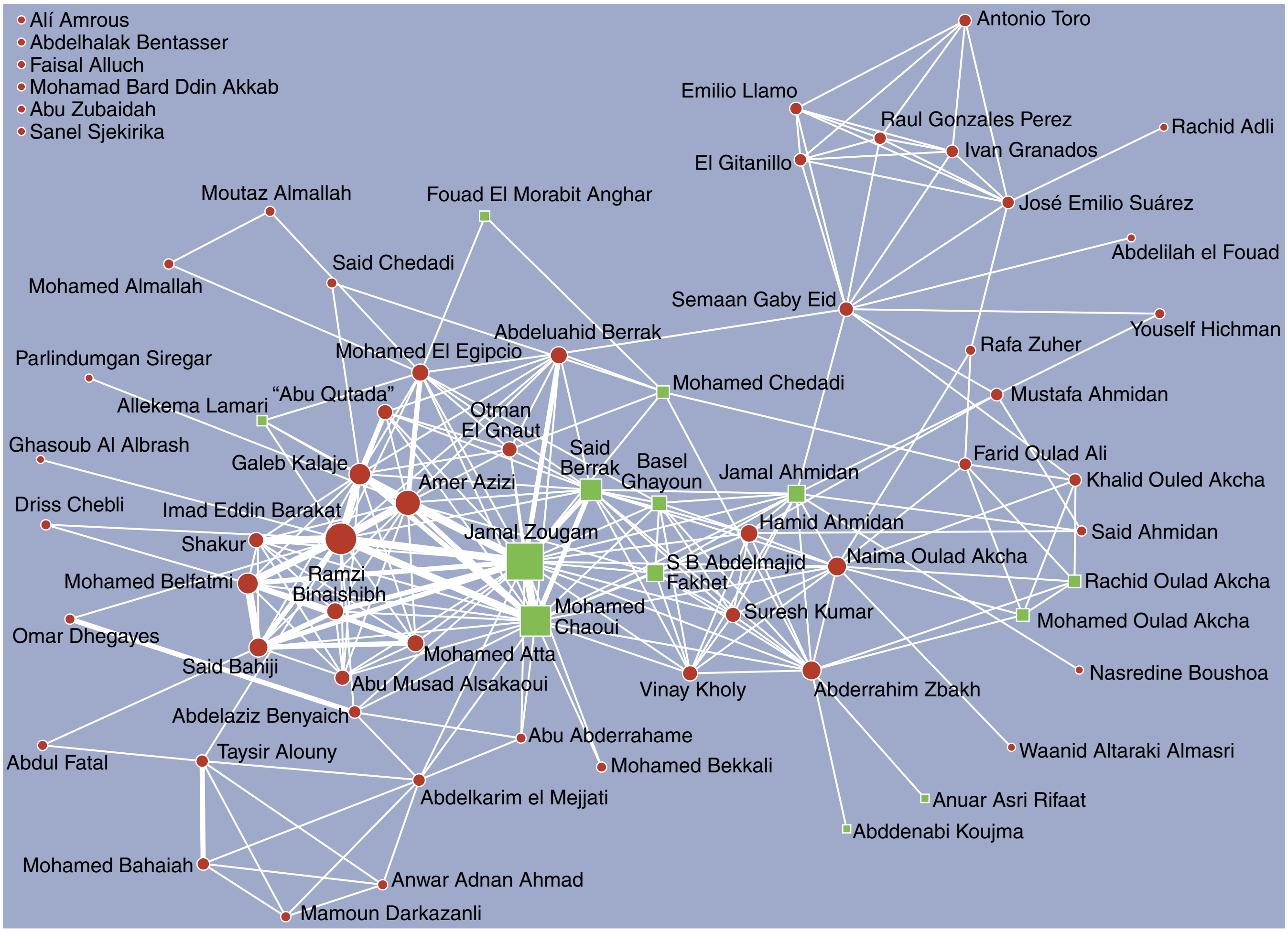
known as traffic analysis. In a battlefield situation, you might intercept an enemy's radio transmissions but be unable to read their encrypted content. Nevertheless, just counting the messages can yield valuable information. A flurry of activity might signal an impending troop movement; sudden radio silence could be even more ominous. If you can identify the source and the intended recipient of each message—in effect, constructing a call graph—you can learn even more, since lines of communication often reveal something about the organization of a military force.

The search for meaningful patterns in telephone records could rely on similar principles, but the problem is much harder. In the military situation, messages between enemy units are readily identified as such. In the telephone database, calls among a few dozen conspirators would all too easily get lost in the background noise of other conversations.

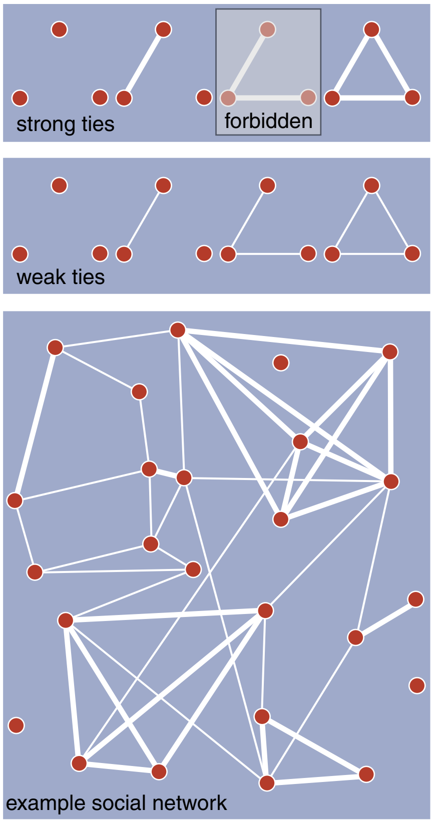
The records in the call database are collected not for the sake of national security but for mundane commercial purposes. In order to send you an itemized bill at the end of the month, a phone company needs to keep track of every call completed, with the originating and receiving phone numbers and the starting and ending times. The largest companies handle roughly 250 million toll calls a day, and so a month's worth of data amounts to several billion call records. AT&T reports that its database of retained records is approaching two trillion calls and more than 300 terabytes of data.

Apart from billing, the call graph has other uses within the phone company—some of which are not too different from what the NSA may be doing, and almost as secretive. Historical calling patterns can be used to detect fraud, and some patterns are also of interest in marketing. For example, a company that offers a discounted rate within a "calling circle" can use information from the call graph to estimate the costs and benefits of the program.

In principle, the same kind of traffic data found in telephone call-detail records could also be compiled for other communications channels. For instance, Federal Express and other courier services keep digitized records of their deliveries, which could readily be transformed into a database of senders and receivers. Curiously, the most digital medium of all—the Internet—does not provide for routine retention of who-speaks-to-whom data; there's no direct need for it, since customers do not pay by the message. However, there is no technological barrier to collecting detailed statistics on e-mail messages and



A map of a social network traces relations among individuals implicated in the bombing of commuter trains in Madrid on March 11, 2004, as well as others thought to be connected with the attack. Green squares represent members of a "field operations group" (which includes those who actually placed the explosives); red circles designate others associated with the group. A white line is drawn between two nodes if the corresponding persons had any of several relationships, such as kinship or frequent presence at a shop owned by two of the conspirators. Thicker lines indicate stronger interpersonal ties. Six persons listed at the upper left are isolated nodes, without any documented links of the kinds examined here. The network was analyzed and the diagram was constructed by José A. Rodríguez of the University of Barcelona.



Triangle rule, proposed in 1973 by Mark S. Granovetter, formalizes the familiar observation that people who have a friend in common are likely also to be friends with each other. Granovetter's model makes this principle a strict rule in the case of strong social ties but not for weaker ones. The rule has a global effect on the structure of social networks: People held together by strong ties must form a *clique*, a complete subnetwork where everyone is linked to everyone else.

other kinds of Internet traffic. A "packet sniffer" installed on the network backbone would simply need to scan the headers of messages and record the *to* and *from* addresses. (It's even possible that equipment reportedly installed by the NSA at certain Internet switching centers could have this purpose.)

Ties That Bind

Digging into the call graph is a form of data mining—and the process could not be more aptly named. Heaving aside hundreds of terabytes of extraneous data is the digital equivalent of a major earthmoving project. Before firing up the steam shovels, it would be helpful to know what we're looking for. What are the communications patterns characteristic of dangerous plotters and connivers?

A good place to turn for an answer to this question is the community of scholars who study social networks: the structures of groups of people as defined by the connections between them. (Of course the community of social-network scholars is itself a social network, held together by many interpersonal connections.)

A seminal paper in this literature, "The strength of weak ties," was published in 1973 by Mark S. Granovetter, now of Stanford University. Granovetter observed that when people are strongly connected—when they are close friends, say, or family members, or working colleagues—the ties between them are usually symmetrical and also obey a rule that might be called triangularity. Symmetry implies that if A is a friend of B, then B is also a friend of A. Triangularity says that if A is friendly with both B and C, then B and C should also be friends with each other. Of course these are merely tendencies, and anyone can cite exceptions (unrequited love, pathological triangles), but for purposes of analysis it's useful to ask what a society would look like if symmetry and triangularity were strictly enforced rules. The answer is that the social structure would consist entirely of perfect *cliques*—groups in which every person is linked to every other person.

Strong bonds between individuals would seem to be the very stuff of social cohesion, but Granovetter's theory suggests a paradoxical effect. Locally, strong ties create highly robust structures, but on a larger scale they also isolate one group from another. Because of the all-or-nothing nature of strong ties, distinct cliques become island universes that cannot communicate with one another. What really holds the world together, Granovetter argues, are *weak ties* between casual acquaintances. These relationships are often symmetrical but seldom triangular: You can chat with a bank teller every week without getting to know all of the teller's other customers. Such weak ties, which at first seem socially insignificant, provide vital cross-links between cliques. The prevailing network structure, according to Granovetter, consists of clusters tightly bound internally by strong ties and loosely linked to other clusters by weak ties.

Social-network theory has obvious affinities with mathematical graph theory—even though people who work in the two fields tend to form distinct clusters linked only by weak ties. Graph theory brings its own vocabulary, as well as a more abstract view of the subject matter: Formally, a graph is a set of vertices together with a set of edges, where each edge connects a pair of vertices. This rather opaque definition can be interpreted in various ways, but in practice graph theorists, like social networkers, draw lots of diagrams with dots and lines.

Plotting the Plotters

What do the principles of social networks and graph theory tell us about the structure of terrorist cells? The very word "cell" offers a clue: It suggests compartmentalization. And indeed the lore of spy rings and resistance fighters speaks of limiting communication so that if one person is captured others will not be put in jeopardy. At the same time, however, the members of the group have to keep in touch in order to make plans and carry them out.

An illuminating case study comes from a rather different context: price-fixing by manufacturers of electrical equipment in the 1950s. The social network of the colluding managers and executives was examined by Wayne E. Baker of the University of Chicago and Robert R. Faulkner of the University of Massachusetts. They found that "the structure of illegal networks is driven primarily by the need to maximize concealment, rather than the need to maximize efficiency." Nevertheless, the price-fixing and bid-rigging simply could not be accomplished without communication among the conspirators, especially in the case of the biggest machinery. Despite the risks, executives had to meet face-to-face to coordinate their plans.

Networks of terrorists apparently face the same conflicting imperatives. Valdis E. Krebs, a consultant who usually applies social-network analysis to business problems, has used the same tools to map relations among the September 11 hijackers. In a paper written just a few weeks after the attacks, he found the network surprisingly sparse. Although every hijacker could be connected to every other via *some* path through the network, many of the paths were quite long, passing through three or four intermediaries. This attenuated structure would make communication extremely inefficient.

Krebs later revised his analysis, as more information became available. He has posted a new map at the Web site http://orgnet.com/prevent.html. Here he reaches a different conclusion. Starting with two men who were already under suspicion in January of 2000, Krebs finds that known linkages lead to all 19 hijackers, and to other conspirators as well. Each node of the network is tied to the two initial subjects either directly or through a single intermediary.

José A. Rodríguez of the University of Barcelona has created a similar network map for the bombing of commuter trains in Madrid on March 11, 2004. Rodríguez recorded several kinds of strong links among the conspirators. Some had ties of kinship or had been childhood friends; others congregated at a shop owned by two of the subjects; some were veterans of earlier wars or terrorist actions. Looking at just the 13 men who actually placed and detonated the explosives, Rodríguez found that the strong ties produced a somewhat strange network. A core of six people formed a clique: Each one was linked to all the others. But the remaining members were only loosely associated or were completely disconnected from the main group.

The outlook changed entirely when Rodríguez included some 70 persons associated with the plot in various ways and when he mapped weak ties as well as strong ones. The weak ties denote pairs of people connected by financial transactions, casual encounters and the like. This larger and fuller network looks much like what Granovetter's theory would predict. There are several dense clusters, within which most nodes are strongly connected, but the clusters communicate with one another only via comparatively loose and unreliable couplings. For example, one cluster is made up of Spanish citizens from whom the bombers obtained explosives; most paths from this subgroup to the rest of the network pass through a single node, a vulnerable choke-point.

Inner Circles and Outer Rings

The social networks described above were constructed retrospectively. The starting point was a complete list of the known members of the group, along with enough biographical information to fill in the links. Discerning the same structure in advance—when an attack is still in the planning stage and most of the plotters are unknown—would be much harder, especially when working from impersonal data such as logs of phone calls or e-mails.

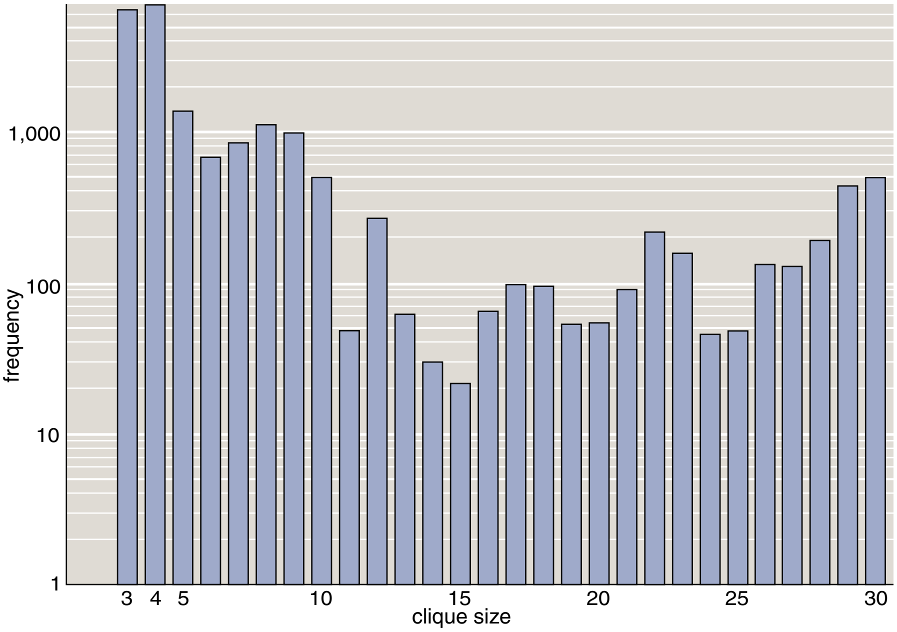
Sketching out such networks surely falls within the NSA's mandate, which might explain the agency's interest in the call graph. One scenario is easy to imagine. Someone has come under suspicion, based on information from other sources, and by consulting the call graph the NSA learns whom that person has been talking with in recent weeks or months. The result is a "ring" of contacts surrounding the subject. Then each of the contacts is investigated in the same way, producing a second ring of contacts-of-contacts. This process could be continued further, although the exponential growth of the graph will soon take in most of the population (especially if the subject has answered a call from a telemarketer or has ordered a pizza). Of more interest are instances where the subject's contacts are also contacts of one another, since such triangular links suggest strong ties.

The tricky part of this network analysis is not finding the links but knowing which of them are significant. It may well be true that if intelligence agencies had "connected the dots," all of the September 11 hijackers could have been linked to the two who were spotted in January of 2000. But thousands of other people would have been linked to those individuals as well. Showing the network of conspirators in isolation is misleading; the graph is actually embedded in a vastly larger structure.

Direct access to the call graph would be a convenience in tracing the associations of known suspects, but it is not necessary. For any named individual, the same records could be obtained under court order, as they are by law-enforcement agencies during criminal investigations. Indeed, if the call graph is used only for such purposes, it seems like quite an extravagance—sifting through 10^{12} records for the few hundred or few thousand calls that might be of interest.

But news reports hint at a more ambitious function for the call graph: not merely tracking down the associates of a known malefactor but rather discovering a plot without any prior guidance, merely by searching the archive for "patterns that might point to suspects." This sounds like magic: You cast your gaze over the vast and intricate web of the call graph, and without even knowing the names behind the phone numbers, you perceive some pattern of linkages that's a danger sign. Can this trick be transferred from the world of magic to the world of algorithms?

If such a signature pattern exists, the findings of social-network theory suggest it should involve some distinctive



Searching for cliques in a massive database of telephone calls could serve as an illustrative proxy for the kind of data mining that intelligence agencies are reported to be attempting. In this context a clique is a set of telephone numbers in which every number was connected to every other number in the set at least once during a given interval. James Abello and his colleagues searched for cliques in a database of more than 170 million phone calls recorded in a single day. Shown here are the number of cliques of each size from 3 through 30 found during one round of the experiment; in a later, more thorough search they found 14,000 cliques of size 30.

combination of strong and weak ties. A terrorist cell, seen from the point of view of telephone traffic, might be a set of people who talk among themselves a lot but have little to say to the rest of the world. Thus the pattern that rings the alarm bells would be a dense subgraph in comparative isolation from its surroundings.

Clique Here

Only the NSA knows whether it can actually spot such patterns, but a somewhat simpler problem can serve as a proxy in estimating the difficulty of the task. The proxy problem is that of finding a large clique within the call graph. This process was studied in the late 1990s by James Abello, then at AT&T Bell Laboratories, and his colleagues.

Finding the largest clique in a graph is a classic hard problem. The brute-force method simply examines every possible subset of vertices and checks to see if all of them are connected. The number of subsets grows so rapidly that this algorithm runs out of oomph even for a graph with 50 vertices; it would be unthinkable for 50 million. The only practical alternatives are approximate and probabilistic methods, which usually converge on a good solution but can't promise to find the best one.

The graph used for Abello's experiments included a single day's records; it had 53,767,087 vertices (corresponding to telephone numbers) and more than 170 million edges (representing calls). The algorithm started with a small clique and tried to build a bigger one. In a first stage, the program repeatedly searched for a new vertex connected with all those already in the set. When no more vertices of this kind could be found, the program switched to another strategy, looking for opportunities to remove one vertex from the clique in exchange for adding two others. Grinding through the day's database took about five hours on a machine with four processors and four gigabytes of memory.

The largest cliques found had 30 vertices. Consider what this means: In a group of 30 telephone numbers, each one either called or was called by all of the other 29 numbers in the course of a single day, for a total of at least 435 calls. That's quite a busy calling circle! But there wasn't just one such clique: Abello's group found more than 14,000 distinct cliques of size 30.

The curious result of this experiment suggests several conclusions, all of them tentative. First, the computational power needed for analyzing call graphs appears to be readily available—though it's not yet on every desktop. Abello's experiments were able to chew through one day's worth of data; today's hardware could doubtless manage much larger bites.

Second, it looks like finding instances of a given pattern within the call graph is not the problem. The problem is defining a pattern selective enough to identify a target group without also branding 14,000 others as possible terrorists. The algorithms must somehow distinguish a few dozen people intent on mayhem from other groups of the same size and structure who are planning a family reunion, canvassing the neighborhood for a lost cat, running for city council or war-dialing to win free concert tickets from a radio station.

No matter what methods are brought to bear on the problem, the intelligence agencies face a formidable task: To survey a huge population (potentially all six billion of us) looking for a tiny subgroup (those planning violence). It's analogous to screening for a rare disease. Even if the test is right 99 percent of the time, almost all the positive results will necessarily be *false* positives.

Abello (who is now with DIMACS, the Center for Computer Science and Discrete Mathematics at Rutgers University, and with Ask.com) thinks that the task of tracing terrorists through call graphs would be difficult, but he also notes that additional information can be extracted from the graphs, apart from the simple pattern matching I have described here. For example, cliques and other clusters have interesting dynamics; some persist from day to day but others vanish. Information like this might help to distinguish one kind of group from another. Abello also mentions extensive recent work on identifying self-organized communities in other contexts, from chat-room participants to eBay customers.

The Plumber's Helper

It's in the nature of secret intelligence programs that most of us will never know for sure what the programs do, how well they work or even whether they exist. Nevertheless, in a democracy citizens can't entirely cede responsibility for what their government may be doing behind the black curtain. To have an informed opinion, we need to puzzle out the facts as best we can. Besides, it's an *interesting* puzzle.

My own opinion, so far, remains ill-formed. Tracking terrorists through call graphs looks like a hard problem. But just because *I'm* stumped certainly doesn't mean it can't be done!

Whether or not call graphs lead to hidden terrorist cells, they may be just the ticket for other tasks. Here's one idea. The Bush administration has expressed displeasure with the public disclosure of all the new surveillance programs, and would like to know who leaked the news. The call graph might be an ideal device for answering that question. One need merely list, on the one hand, all those who had access to the information, and on the other hand the journalists who ultimately reported the story. Search in the graph for direct or indirect connections between those two sets of vertices. The irony is that whoever released the information probably understood quite clearly this potential for exposure.

Bibliography

Abello, J., P. M. Pardalos and M. G. C. Resende. 1999. On maximum clique problems in very large graphs. In *External Memory Algorithms*, edited by James M. Abello and Jeffrey Scott Vitter, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 50, pp. 119–130. Providence, R.I.: American Mathematical Society.

Abello, James, Mauricio G. C. Resende and Sandra Sudarsky. 2002. Massive quasi-clique detection. In *LATIN 2002: Latin American Theoretical Informatics Symposium*. Lecture Notes in Computer Science vol. 2286, pp. 598–612. Berlin: Springer Verlag.

Baker, W. E., and R. R. Faulkner. 1993. The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. *American Sociological Review* 58(6):837–860.

Chakrabarti, Deepayan, and Christos Faloutsos. 2006. Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys* 38:1–69.

Granovetter, Mark S. 1973. The strength of weak ties. *American Journal of Sociology* 78:1360–1380.

Hayes, Brian. 2000. Graph theory in practice: Part I. *American Scientist* 88:9–13.

Hayes, Brian. 2006. Room 641A. http://bit-player.org/2006/room-641a

Krebs, Valdis E. 2001. Mapping networks of terrorist cells. *Connections* 24(3):43–52. http://www.sfu.ca/~insna/Connections-Web/Volume24-3/Valdis.Krebs.web.pdf

Krebs, Valdis. Web site. Connecting the dots—tracking two identified terrorists. http://orgnet.com/prevent.html

Rodríguez, José A. 2005. The March 11th terrorist network: In its weakness lies its strength. Working paper EPP-LEA no. 3, Grupo de Estudios de Poder y Privilegio, Departament de Sociologia i Anàlisi de les Organitzacions, Universitat de Barcelona. http://www.ub.es/epp/wp/11m.PDF