Mr. Ted Senator
Information Awareness Office (IAO)
Evidence Extraction and Link Discovery Program

Good morning. I'd like to take this time to describe our Evidence Extraction and Link Discovery Program—EELD for short. I'll begin by explaining and characterizing the problem we're trying to solve, and follow by summarizing the approaches we are pursuing and some of our initial successes.

I'll start off by asking everyone an easy multiple-choice question: What is the main reason that you came to DARPATech? Was it a) to learn about what DARPA is doing? b) to make connections with DARPA program managers or with potential business partners? c) to try to get intelligence about what your competitors are up to? or d) some other reason?

OK, let's have a show of hands.

How many of you came to learn about DARPA's programs and plans?

How many to make connections?

How many to gather competitive intelligence?

And how many for some other reason?

Well, it appears to me that most of you came to make connections.

Let me ask another question. How many transactions did you make to arrange to attend?
Let's do another show of hands. None?

OK, (almost) nobody picked none.

How about one or two? A few of you picked that.

How about three to five ? Six to ten?

More than ten?

I personally made more than 25 transactions.
I had to arrange for airline tickets and hotel reservations and airport transportation.
I sent e-mails to colleagues and to friends to coordinate schedules.
I had to coordinate schedules with my wife and children.
I checked airline reservation web sites for flight options.
I registered.
And I must have sent and received innumerable e-mails with various drafts of this talk.

All this adds up to quite a large number compared to what most of you chose. And I bet my experience really wasn't atypical. Most of us make many more transactions than we realize.

Let's try one more question. How far in advance did you begin planning for DARPATech?

When the dates were announced? When registration was opened? Before then? After?

After the deadline for advance-purchase discount airline tickets?

Those of you who decided or made arrangements earlier, I bet you are from out of town. And I bet that those of you who made fewer transactions and did them later are closer or even local to southern California. There

may even be some local attendees who decided at the last minute to participate (at least as close to the last minute as our security system permits).

The points I am trying to illustrate with this example are that not only do people make many more transactions than we realize, but the more complex the plan, the farther in advance these transactions begin. These transactions fit together in patterns, which can be recognized, with increasing certainty as we get closer to the conference. If I had access to your transactions, and of other attendees, and perhaps of other travelers to this area for the past few days, or you mine, I could tell if you are traveling alone or with a colleague or with your family. I could tell if you are the type of person who can make advance plans and buy tickets in advance or the type who needs the flexibility to change at the last minute. If I combined registration information I could learn a lot about the demographics of attendance. I could identify DARPA personnel, other Government personnel, and representatives of large and small contractors. I could also find people who fit cleanly into a category, or who don't. With enough information I could perhaps distinguish between people visiting Disneyland for a vacation and those coming to DARPATech. And, very importantly, I could do this without identifying you, as long as I could associate all your transactions with a single individual and preserve the links between individuals who appear in the same transactions.

Now imagine that I didn't know anything about DARPATech. How could I have inferred its existence by looking at all the transactional data for people traveling to Southern California? Perhaps I would have noticed a large group of people coming to a meeting at this hotel. But this hotel has conventions all the time, so there might not have been anything out of the ordinary. What might have tipped me off? What makes this meeting different from many others at this hotel? I'll let you think about this one.

Let's generalize from this example. Detecting interesting patterns of activity consists of finding connections between people, organizations, places, and things, and then recognizing interesting patterns of activity conducted by these people and organizations. These connections may be manifested as transactions between people and or organizations, by common membership in organizations, or by frequent co-occurrence of the same people in multiple events. We've all seen what's meant by links and relationships in the past year. Many newspaper articles have appeared about the events of September 11, typically accompanied by very nice graphics that show the relationships between the hijackers—some roomed together in Hamburg, some had airline tickets purchased on the same credit card at the same time, some traveled to Las Vegas at the same time, and the pilots trained together and, most important to our ability to have detected the plot in advance, engaged in suspicious and unexplainable behavior that was reported during this training. These articles had as their theme: "we had the information but didn't put it together."

And that is what EELD is all about: developing techniques that allow us to find relevant information—about links between people, organizations, places, and things—from the masses of available data, putting it together by connecting these bits of information into patterns that be evaluated and analyzed, and learning what patterns discriminate between legitimate and suspicious behavior.

This is not an easy task. But it is one about which we have some new ideas and have been making significant progress. Before I tell you about these ideas and this progress, I'll first explain why existing techniques are not up to the task.

Traditional fraud detection techniques look for outliers, i.e., behavior by individuals that is unusual according to some statistical measure. It may be unusual because it differs from a peer group, e.g., a cashier with a much higher loss ratio than typical, or it may be unusual because it represents a major change by the same individual, e.g., a change in a cellular telephone calling pattern or a credit card spending pattern. Models used in these fraud detection techniques can be developed through techniques such as statistical time series analysis, data mining, machine learning, or even neural networks, and are well developed in industry and Government. They assume that we know an individual or account holder's identity, or, at the very least, that we can uniquely identify individuals or accounts and combine information about the individual or account to create a profile of that individual or account's typical behavior. Metaphorically, these techniques aim to find a needle in a haystack, or viewed another way, to construct a jigsaw puzzle without the picture on the box.

Contrast this type of fraud detection with what must occur for asymmetric threat detection. What we need to look for to detect behavior patterns representative of asymmetric threats is not outliers, but what I like to call "in-liers." Activities such as getting a pilot's license, or purchasing airline tickets with cash at the last minute, or overstaying a visa, are not by themselves indicators of terrorist behavior. In fact, the most dangerous adversaries will be the ones who most successfully disguise their individual transactions to appear normal, reasonable, and legitimate. It is only when they are combined in a particular way, or, in other words, they occur in a particular context, that they become highly suspicious.

Using the first of the previous two metaphors, our task is akin to finding dangerous groups of needles hidden in stacks of needle pieces. This is much harder than simply finding needles in a haystack: we have to search through many stacks, not just one; we do not have a contrast between shiny, hard needles and dull, fragile hay; we have many ways of putting the pieces together into individual needles and the needles into groups of needles; and we can not tell if a needle or group is dangerous until it is at least partially assembled. So, in principle at least, we must track all the needle pieces all of the time and consider all possible combinations.

Using the second of the two metaphors, our task is more akin to having millions of jigsaw puzzles without the pictures, and having access to only a small fraction of the pieces. This metaphor illustrates why simply having more analysts can not solve the problem—the likelihood that any two, let alone more than two, pieces are from the same puzzle goes down as the set of pieces is divided up among more analysts. It also illustrates why current practice is insufficient—one must know how to divide up the problem in a way that maximizes within-group connections and minimizes cross-group connections—for this approach to succeed. As we all know, this separability could be achieved in the era of symmetric threats, when threats corresponded to countries and alliances, but it can not be achieved in the current era of transnational threats comprising small groups of individuals.

How can we get visibility into a network that relies on a small group of individuals with strong, trusted relationships? The best way, according to Valdis Krebs, a leading student of social networks, "may be to discover possible suspects and then, via snowball sampling, map their individual personal networks—see whom else they lead to and where they overlap." This is the approach that guides the technology development for EELD.

How do we get a starting point? There are three possible approaches. One is to match up large transaction databases, looking for unusual spatio-temporal co-occurrences. A second approach is to monitor data streams for known or suspected indicators of illicit activity. And another is to take advantage of what we already know—to use effectively all of the large amount of law enforcement and intelligence information that we already collect.

Much technology exists to implement the first approach. While it can find groups of people who appear to be linked together, it tells us nothing about whether their activities are legitimate or suspicious. It also requires us to make many assumptions about the prior joint probability distributions, such as the likelihood that two people will happen to be at a particular airport together. And there are important and legitimate legal and policy constraints that prohibit its widespread use.

Monitoring data streams for indicators of activity can suffer from the limitations of traditional fraud detection techniques and also can be limited to finding instances of previously known or suspected types of threatening behavior. Criteria for new types of threatening behavior can be incorporated after they are discovered, typically after a small but significant number of instances of that behavior surface. This reaction time, of allowing for a small number of new types of incidents before updating the automated system, may work for domains where the goal is preventing most illicit behavior most of the time, such as credit card fraud, but is not acceptable for the one-time rare events of such magnitude that we experienced on September 11.

EELD is developing technology for the third approach—to extract from documents such as law enforcement records or other intelligence reports relational facts that provide evidence of connections between entities, and to store these facts as links in an evidence database. While this is a large amount of information, it is finite and it is small compared to all the transactional and textual information in the world. We then use these

as starting points to gather additional information—from textual and transactional sources—about the individuals identified in these reports and their associates, all the while evaluating the significance of the emerging networks of relationships and the activities conducted by these individuals—to guide our search. We look for evidence of known patterns and, perhaps more important, for unexplainable connections that may indicate previously unknown but significant connections, representing, for example, a new group, threat, or capability.

What ties all this together? It is the idea that detecting patterns of activity, which comprise links between people, organizations, places, and things, requires extending technology in three key areas to handle structured data representations. These three areas are evidence extraction, link discovery, and pattern learning. We imagine a large database of evidence, represented as a graph, with nodes signifying entities and links signifying relationships. We need to populate this database using evidence extraction techniques. We need to connect fragments of evidence into meaningful patterns using link discovery techniques. And we need to learn new patterns using pattern learning techniques. Let me tell you about some of the technology we are developing.

Our starting point is considered to be intelligence reports. Recall all the newspaper stories about how we had reports of suspicious flight training activities prior to September 11. We are extending information extraction technology to be able to extract relationships, or links, between entities. Note that we are not attempting to solve the full extraction problem of scenario-based event extraction, in which one attempts to describe all aspects of a complex event, such as the multi-year timeline of activities pursued by the hijackers. Rather, we are extending current technology which can extract entities and attributes to the ability to extract relationships and their attributes, including the participating entities and their roles. We hope to achieve the same 90-95% accuracy on relational facts that is currently achievable with the best technology for entities and attributes. We are also attempting to achieve rapid portability to new domains. And we are working closely with Charles Wayne and his TIDES program to build on the many advances in natural language processing he is achieving.

We are following three alternative approaches at BBN, SRA International, and Syracuse University. BBN is using an integrated statistical language model, which is trained on general linguistic knowledge and smaller specialized domain-specific training sets. Syracuse University's Center for Natural Language Processing is using Transformation-Based Learning to learn automatically the extraction rules for new domains. SRA is extending their commercial NetOwl extraction system with an ontology of link types of interest to the intelligence community and exploring genetic programming techniques for rapid adaptability to new domains.

Once we have a database of evidence constructed from our extractors, we need to discover unknown links and build up the set of connections. We have multiple approaches to this problem too. At Carnegie-Mellon University, Andrew Moore and Jeff Schneider have extended the idea of merging of databases to simultaneously learn optimal models of link probabilities, group membership, and demographic classification models based on demographic (i.e., attributes) and link data about individuals. These techniques provide a formal and scalable approach to database matching that can be used to estimate the likelihood of meaningful linkages between individuals based on their observed transactions. It can also be applied to determine the likelihood that multiple aliases refer to the same individual.

Doug Lenat and his colleagues at Cycorp are adding ontological concepts and knowledge to the Cyc Knowledge Base to enable richer link discovery. For example, it will be able to determine using travel time constraints if the same alias in two different places can really be the same individual, or to determine whether there is a relationship between a particular group and a particular individual, or whether one individual trusts or has influence over another. Consistent with the EELD vision of constructing patterns of activity from connected observations, the answers to these queries are not assertions but graphs representing relationships.

ALPHATECH is using a combination of data fusion and AI techniques to build an integrated link discovery component. They are developing a multi-hypothesis Bayesian reasoning system for linking and updating

attributes of relational data and combining it with deterministic logic and constraint-based reasoning methods to manage efficiently the number of hypotheses to generate and consider.

SRI International is developing a partial pattern matcher based on a graph-edit distance metric that will enable the expression and matching of patterns expressing a rich set of relationships—hierarchical, temporal, uncertain, qualitative, and uncertain.

USC's Information Sciences Institute is developing a set of knowledge-based and statistical tools, including a pattern instantiator, connection finder, link elaborator, partial matcher, cluster engine, phase transition monitor, object identifier, and group hypothesizer/ checker. Also at ISI, Craig Knoblock is extending his WIDELink system to enable the rapid collection of link information from web sites. Building on existing search engine technology, WIDELink will allow us to populate databases with relational facts from web pages.

Metron is extending likelihood ratio-based non-linear tracking techniques to recognize instances of possible execution of terrorist plans. 21st Century Technology is applying graph-theory based techniques to efficiently match previously defined structures in a large database. And CHI Systems is using cognitive models and case-based reasoning to detect potential terrorist plans as instantiations of a set of invariant mission-planning templates.

Finally, we are exploring techniques of pattern learning to allow us to identify previously unknown patterns of activity, which would be used to guide the link discovery process.

Jude Shavlik and his collaborators at the Universities of Wisconsin and Texas are extending and applying inductive logic programming techniques. ILP naturally represents relational data; challenges are scalability and incorporation of uncertainty. Daphne Koller of Stanford has extended her prior work on probabilistic relational models, allowing for undirected graphs, and achieving marked improvements in classification tasks. Foster Provost at NYU and David Jensen at the University of Massachusetts are working to extend traditional knowledge discovery and data mining techniques to relational data. Paul Cohen, also at University of Massachusetts, is exploring a variety of techniques to recognizing temporal patterns. And Larry Holder and Diane Cook at the University of Texas at Arlington are developing graph reduction techniques that identify and collapse common substructures, enabling more effective pattern recognition.

In conclusion, let me once again restate our hypothesis by reminding you of the Defense Science Board's study on Transnational Threats. They stated: “the making of connections between otherwise meaningless bits of information is at the core of (transnational) threat analysis" and "Search methods currently in use are not up to the challenge." EELD is working to extend key technologies in three areas—evidence extraction, link discovery, and pattern learning—to enable the making of connections between otherwise meaningless bits of information, by constructing an evidence database of relational facts, by applying a combination of knowledge and graph theory based techniques for efficient searching and pattern recognition, and by enabling the learning of previously unknown structural patterns of interest.

We are making significant progress on some key technical issues. We will be integrating promising techniques into the prototype TIA system to evaluate their effectiveness not only in the laboratory but also on real problems with real data. We will make it possible to get the necessary information and to put it together.

Thank you.