Using Importance Flooding to Identify Interesting Networks of Criminal Activity

Byron Marshall
Accounting, Finance, and Information Management Department, Oregon State University, Covallis, OR 97331. E-mail: byron.marshall@bus.oregonstate.edu

Hsinchun Chen
Department of Management Information Systems, University of Arizona, Tucson, AZ 85721. E-mail: hchen@eller.arizona.edu

Siddharth Kaza
Department of Computer and Information Sciences, Towson University, 8000 York Road, Towson, MD 21252. E-mail: skaza@towson.edu

Effectively harnessing available data to support homeland-security-related applications is a major focus in the emerging science of intelligence and security informatics (ISI). Many studies have focused on criminal-network analysis as a major challenge within the ISI domain. Though various methodologies have been proposed, none have been tested for usefulness in creating link charts. This study compares manually created link charts to suggestions made by the proposed importance-flooding algorithm. Mirroring manual investigational processes, our iterative computation employs association-strength metrics, incorporates path-based node importance heuristics, allows for case-specific notions of importance, and adjusts based on the accuracy of previous suggestions. Interesting items are identified by leveraging both node attributes and network structure in a single computation. Our data set was systematically constructed from heterogeneous sources and omits many privacy-sensitive data elements such as case narratives and phone numbers. The flooding algorithm improved on both manual and link-weight-only computations, and our results suggest that the approach is robust across different interpretations of the user-provided heuristics. This study demonstrates an interesting methodology for including user-provided heuristics in network-based analysis, and can help guide the development of ISI-related analysis tools.

Introduction

The growing science of intelligence and security informatics (ISI) explores the use of advanced information technology in national/international and homeland-security-related applications. A key underlying problem is the diversity and volume of information that needs to be disseminated, analyzed, and acted upon (Raghu, Ramesh, & Whinston, 2005). Learning to extract useful leads from law enforcement data is both specifically important for homeland security processes (local, regional, national, and international) and more generally important as an exemplar for complex analysis tasks that deal with ambiguous relationships, suffer from missing information, employ user-provided heuristics , and are usefully represented as networks.

Although a number of national and regional data-sharing systems have been deployed (with varying degrees of success) there is little agreement on which data should be shared or how data should be analyzed to support investigations. The Global Justice XML Data Model (GJXDM) supported by the U.S. Department of Justice (DoJ) Office of Justice Programs (OJP; http://www.it.ojp.gov/jxdm/), which specifies entities and attributes appropriate for encoding and sharing criminal-justice data, is a start, but computer-supported investigational models are needed to guide the development of investigationally useful policies, protocols, and procedures. This work aims to develop an analysis methodology that employs shareable data (minimizing privacy, formatting, and administrative concerns) to address a real-world investigational task (link-chart creation), incorporating the kinds of heuristics employed by investigators.

Network-based techniques are widely used in criminal investigations because patterns of association are actionable and understandable. Investigators identify a suspect's known associates to generate leads and to obtain criminal conspiracy convictions that can keep dangerous criminals off the street

Received August 14, 2007; revised June 2, 2008; accepted June 2, 2008

 • Published online 16 July 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20924

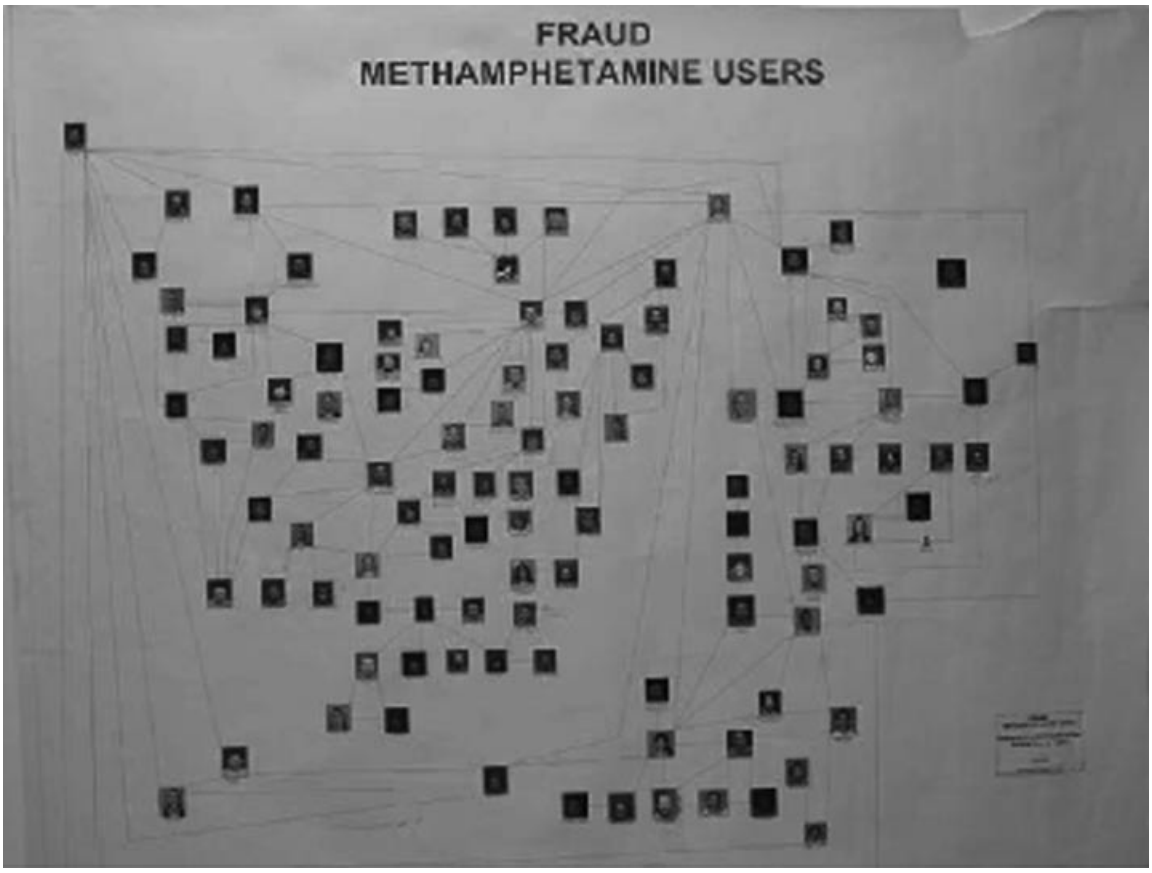

FIG. 1. It took an experienced crime analyst six weeks to extract this Fraud/Meth Link Chart from Tucson Police Department and Pima County Sheriff's Department records. Beginning with a few suspects, a network of associations was identified to help with investigations involving fraud and methamphetamine trafficking.

for longer periods of time. Although many association networks are "drawn" only in the minds of the investigators, visual network depictions called *link charts* are commonly used in important cases. They combine multiple events (based on crime types, localities, or target individuals) to depict a focused set of criminal activity. Link charts help focus investigations, communicate within law enforcement agencies, and present data in court. Figure 1 is a link chart created for the Fraud Unit of the Tucson Police Department. Even though most of the information in the chart came from computerized records, it took an experienced crime analyst several weeks to create the chart. Tools to help this process along are desirable because link-chart creation is also representative of other common investigational tasks.

This article addresses the following research question: How can we effectively and more efficiently identify useful associations for link-chart creation in large collections of criminal incidents, employing investigation-specific heuristics to generate leads and support criminal conspiracy investigations? Expanding on previously published work (Marshall & Chen, 2006) we develop an enhanced link-chart creation methodology to (a) save time and money, (b) allow the technique to be used in more investigations, and (c) automatically employ large quantities of available data. Such a model could be used to support investigations and guide the implementation of data-sharing systems. We report on experiments that compare a range of manual and semiautomated approaches. In the Literature Review, we discuss previous relevant research. We then present our system design and describe our importance-flooding algorithm, which uses spreading activation, path-based importance heuristics, and a learning component that adjusts based on the accuracy of previous suggestions. The next section describes our data set and test cases, and the section following presents our experimental design. These experiments both test the accuracy of our suggestions and explore the sensitivity of our computations to expected heuristic and parameter variations. Our results are reported, then discussed, and we end by identifying future directions.

Literature Review

The methodology presented in this article extracts interesting subsets from cross-jurisdictional data sets using network-based techniques, building on the previous criminal-network analysis work described below. Previous work in data mining explores the connection between network-based analysis and interestingness as discussed in the subsequent sections. Investigational applications of criminal-activity network analysis are imprecise in that the meaning of the links is ambiguous,

DOI: 10.1002/asi

important data is missing, and decision rules are subject to interpretation. We conclude the literature review with a description of the similarity-flooding algorithm, which can leverage a variety of heuristic input rules as it computes node-pair similarity in ambiguously represented networks of relationships.

Criminal-Network Analysis

Criminal-activity networks are frequently analyzed using manually produced link charts. Link charts (e.g., Figure 1) have been used for several decades in the law enforcement domain (Coady, 1985; Coffman, Greenblatt, & Marcus, 2004; Klerks, 2001) depicting individuals and relationships discovered in the course of an investigation. Most of the related research focuses on assigning roles to actors in a network. Sparrow (1991) explored social-network measures (e.g., centrality) as they apply to criminal networks. He points out that questions such as " 'Who is central to the organization?,' 'Which names in this database appear to be aliases?,' 'Which three individuals' removal or incapacitation would sever this drug-supply network?,' 'What role or roles does a specific individual appear to play within a criminal organization?' and 'Which communications links within a international terrorist fraternity are likely to be most worth monitoring?' " (p. 252) would all be familiar to social-network analysis (SNA) practitioners.

Some of the analysis techniques anticipated by Sparrow have been explored in more recent work. Krebs (2001) used centrality measures to identify the group leader of the September 11th hijackers. Another terrorist-network study calculated the average degree of the Jemaah Islamiyah terrorist network (Koschade, 2006) and uncovered that the 2002 Bali bombing cell had a high density that allowed it to sustain member losses. Xu and Chen (2004, 2005) used SNA methods to determine the leader and gatekeeper role for individual nodes, and used hierarchical clustering methods to identify subgroups in criminal networks. Kaza, Xu, Marshall, and Chen (2005) explored the topological characteristics of cross-jurisdictional criminal networks and later studied (Kaza, Hu, & Chen, 2007) link formation and evolution processes. Some of the above techniques have been implemented in crime-analysis tools.

First-generation tools take a manual approach, allowing investigators to depict criminal activity as a network of associations. Second-generation systems include Netmap (Chabrow, 2002), Analyst's Notebook (I2, 2004), and the COPLINK Visualizer (Chen, Zeng, Atabakhsh, Wyzga, & Schroeder, 2003). These tools provide various levels of interaction and pattern identification, representing information using various visual clues and algorithms to help the user understand charted relationships. Third-generation tools possess advanced analytical capabilities. This class of tools has yet to be widely deployed, but techniques and methodologies have been explored in the research literature. Coffman et al. (2004) introduces genetic algorithms to implement subgraph isomorphism and classification via social-network-analysis metrics for intelligence analysis. Network-analysis tools to measure centrality, detect subgroups, and identify interaction patterns were used in Xu and Chen (2003). Most of the above tools identify key nodes and links using centrality and other topological measures. They do not incorporate interestingness based on the semantics of the nodes and relationships between them.

Interestingness Measures

Notions of interestingness have received special attention in the context of data that can be represented as a network. In a law enforcement context, shortest-path measures have been applied to the task of identifying an individual's closest associates. CrimeLink Explorer employed relation-strength heuristics to support shortest-path analysis (Schroeder, Xu, & Chen, 2003). Based on conversations with domain experts, they weighted associations by (a) crime type and person-role, (b) shared addresses or phones, and (c) incident co-occurrence. An algorithm for shortest-path analysis for criminal networks was implemented and tested in Xu and Chen (2004). Because criminal networks can be very large and very dense, the computational burden required to identify the shortest path between two individuals can be significant. Xu and Chen (2004) address this using a carefully crafted computational strategy.

Integrating association-rule mining and criminal-network-analysis tools may help better support analysts as they try to identify "interesting" subsets of large criminal-activity networks. Interestingness measures assign a ranking to discovered associations based on some interestingness calculation methodology (Hilderman & Hamilton, 2001). These measures can be categorized as being either objective or subjective (Silberschatz & Tuzhilin, 1996). Objective measures are generally statistical and include confidence and support. Subjective measures can be classified into two groups: actionable and unexpected. Padmanabhan and Tuzhilin (1999) note that beliefs are important in identifying interesting associations. Results can be filtered by encoding user beliefs (e.g., expected or potentially actionable relationships or patterns) using some "grammar," and comparing extracted relationships to that grammar (Sahar, 2002). A way to incorporate beliefs is important for automatic interestingness analysis.

Network-Based Interestingness

Some researchers emphasize interestingness as a network-based phenomenon. For example, starting from a "root set" of nodes can enhance relevance searching. S. White and Smyth (2003) describe a class of algorithms that incorporate explicit definitions of relative importance in a network context. Based on a scalar coefficient, smaller amounts of importance are passed as distance increases. The two main intuitions behind the approach are that (a) two nodes are related according to the paths that connect them, and (b) the longer a path is, the less importance is conferred along that path. These notions of relative importance align well with the cognitive

DOI: 10.1002/asi

model described by investigators who begin with some target suspect(s) and look for close associates to identify leads.

Previous studies (Xu & Chen, 2003; H.D. White, 2003) employ network path information to identify interesting nodes and links. However, in Lin and Chalupsky (2003), novel network paths (not just nodes or links) are identified to reveal interesting information. Bibliographic citation data from the Open Task of the 2003 KDD Cup (Gehrke, Ginsparg, & Ginsparg, 2003) was analyzed to answer questions such as "Which people are interestingly connected to C.N. Pope?" The basic notion of their analysis was to detect interesting short paths through a network rather than to detect interesting nodes. They categorized link types and used multiple node types in their network. So, for instance, universities were associated with authors who had published a paper while affiliated with the university, and authors were associated with their coauthors. Without putting in specific rules defining *interesting*, their algorithm discovered that Mr. H. Lu was the most interesting person relative to C.N. Pope because he interacted with Pope along a variety of network paths. These paths take the following form:

> [Lu]-writes-[Paper1]-cites-[Paper2]-written_by-[Pope]
> [Lu]-authors-[Paper1]-authored_by-[Pope], and
> [Lu]-authors-[Paper1]-authored_by-[Person1]-authors-[Paper2]-authored_by-[Pope].

This notion that interestingness is path-based rather than node-based is applicable to criminal investigations. For example, one analyst working on a Fraud/Meth link chart noted that she was most interested in people who sold drugs and were associated both with people who sold methamphetamines and people who committed fraud. This kind of association pattern can be viewed as a short path through the criminal-activity network.

Network-based interestingness computations are somewhat similar to relevance computations that leverage structured lexical and semantic resources. The ANSI/NISO Z39.19-2005 standard notes that development of controlled vocabularies (one type of semantic resource) is intended to improve information-retrieval effectiveness (NISO, 2005). The primary purpose of vocabulary control is to achieve consistency in the description of content objects and to facilitate retrieval. Z39.19 describes an appropriate process for defining the relationships between terms, giving careful thought to ambiguity, synonymy, relationship identification, and validation. When retrieval systems leverage controlled vocabularies for searches (notably through query expansion), different types of relationships add different types of information to the analysis. For example, Greenberg (2001a) explores the relative value of various types of term relationships, such as narrower terms, broader terms, synonyms, and related terms. In another article, partial synonyms and narrower terms are identified as most useful in automatic query expansion (Greenberg, 2001b), while related terms are shown to be better candidates for interactive query expansion. This line of work builds on many previous studies that utilize lexical resources such as WordNet (Miller, 1995) to enhance retrieval. Relationship types in a thesaurus reveal the structure of the vocabulary, and the likelihood that a particular class of relationship will increase retrieval accuracy can be assessed. In Greenberg (2001b, pp. 487), the relative value of different classes of relationship type was consistent "regardless of end-users' retrieval goals." Like relationships found in a controlled vocabulary, relationships inferred from law enforcement records can be classified by type. However, while the value of some relationship types (e.g., shared addresses or incident co-occurrence) has been explored as cited above, it is not yet clear how the relevance-inferring power of different relationship types will vary over different investigational tasks.

Similarity Flooding

The similarity-flooding algorithm (Melnik, Garcia-Molina, & Rahm, 2002) was designed to support schema matching using an iterative, network-based computation to overcome link and label ambiguity in matching ontologies and database schemas. The algorithm builds a pairwise connectivity graph of nodes connected by edges derived from the typed links found in two comparable graphs. Each node represents a pair of entities, one from each graph. Similarity scores for the nodes are repeatedly adjusted based on values passed along the edges of that network using a fixpoint computation. The computation terminates when the node similarity scores stabilize. In the original tests of the algorithm, pairs of database schemas are represented as networks of labeled nodes with links based on attributes such as data type and foreign key indicators. Initial similarity is established using string-match heuristics and "flooded" through the network of structural relationships. The algorithm suggests possible matches between elements in the schemas for manual evaluation to save the user time in aligning two models or schemas. It has also been applied in evaluating student-drawn concept maps (Marshall & Chen, 2006) to help analyze similarly ambiguous networks. The technique allows multiple characteristics to contribute to its suggestions, and allows similarity to be passed to potential matches through several network hops. The notion of heuristic rules, applied over multiple hops of weighted associations, to create promising suggestions for manual evaluation is appropriate for application in law enforcement analysis tasks. The basic intuition of the similarity-flooding approach is that a pair of nodes is more likely to be similar when it is connected to pairs of similar nodes. In a law enforcement context we would say a person who is closely associated with an important person is more likely to be important.

System Design and the Importance-Flooding Algorithm

Recognizing that criminal records can be usefully organized into networks of associations, we designed the system depicted in Figure 2. Based on our review of the literature and

DOI: 10.1002/asi

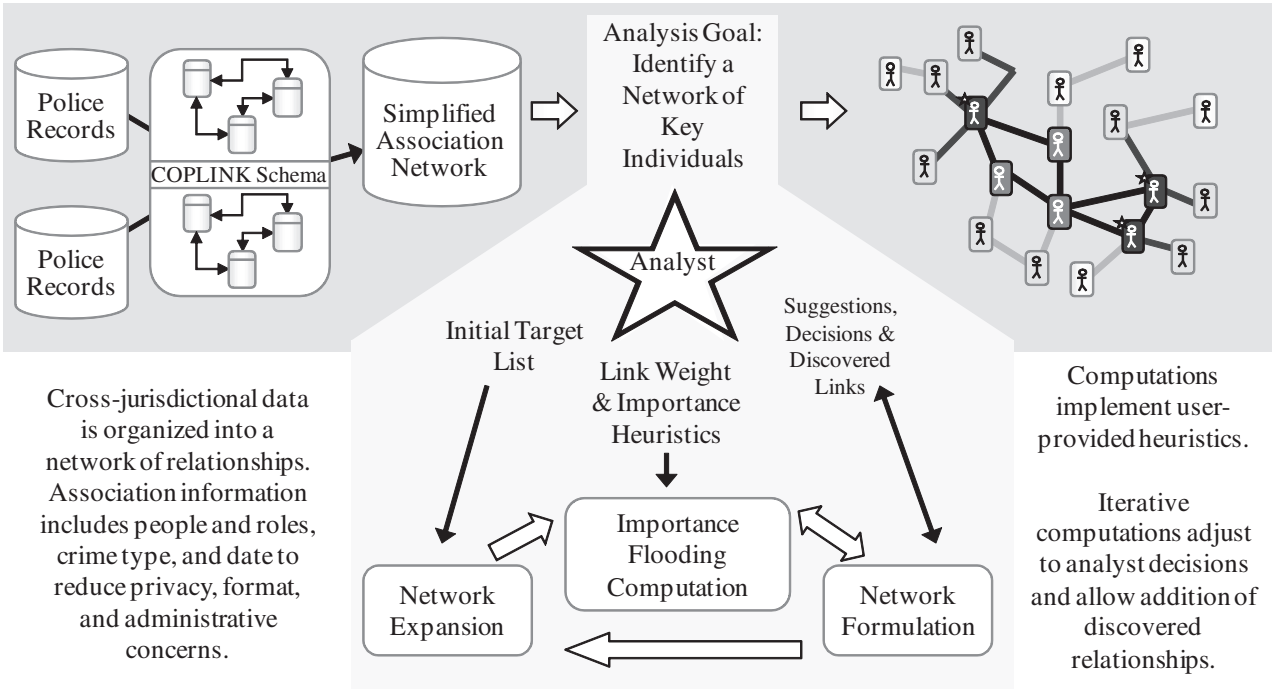

FIG. 2. Relations are extracted and formatted to form a large criminal-activity network, the importance-flooding algorithm iteratively suggests potentially interesting individuals as leads, and the analyst constructs a link chart. In our experiments, data was extracted from the COPLINK systems of two police agencies.

input from investigators, we developed several system-design goals.

1. Use shareable data. Although our methodology can employ a wide range of inputs, the experiments presented in this article use data that has been carefully selected to avoid many of the privacy and security limitations applicable to law enforcement analysis. The Global Justice XML Data Model (GJXDM) allows for entities and relationships so that generated link charts can be shared using this standard.
2. Leverage investigation-specific information not included in the main data set. As analysts review previous incidents, they often come across additional, potentially interesting relationships, for example, family members. Our system can use additional links added as the analysis proceeds. We test the usefulness of this kind of data in the "discovered-link" treatment described in the Results section.
3. Be target focused. Our methodology begins with an investigational target or targets rather than "fishing" through records for general patterns of activity in all of our experimental treatments. We test the sensitivity of the approach to different yet reasonable sets of target individuals as reported in the Sensitivity Analysis section of the Results.
4. Incorporate domain- or investigation-appropriate heuristics (or beliefs) to support analysis, encoding these heuristics in a format that can be adjusted at query time for new insights. We use these beliefs in several ways in our methodology: (a) General association heuristics are used to establish link weights; (b) using initial importance rules, an analyst expresses beliefs about what makes an individual interesting; and (c) rules can be path-based in that they recognize general patterns of behavior that signal importance. Our experiments test different formulations of these importance rules and the sensitivity of our results to reasonable variations in how the heuristics are expressed, as reported in the Sensitivity Analysis section of the Results.
5. Tolerate missing and ambiguous data. Missing information is expected to hamper analysis, but good methodologies need to be tolerant of data limitations. The flooding approach used in our methodology allows importance to transitively pass amongst individuals, compensating somewhat for occasional missing links.

Importantly, these goals are applicable to both smaller, local and large-scale cross-jurisdictional investigations.

System Overview

The underlying process model proceeds in three steps as depicted in Figure 2:

1. Organize cross-jurisdictional data in a simplified association network;
2. Analyze the network to identify interesting associates of the target; and
3. Suggest individuals for possible inclusion in the link-chart creation.

Numerous relationships can be extracted from police records to form an association network. When two people are listed in an incident report, an association between them is inferred. Previous research aimed at measuring association strength has used a variety of indicators to assess the

DOI: 10.1002/asi

strength of the relationship between two individuals. While our methodology allows us to include nodes such as cars, weapons, and addresses, our current network-building process uses only person-to-person connections from police incidents to build the base network.

The analysis module computes an importance score for an individual in the network using heuristically established link weights and initial importance values. This approach imitates how detectives evaluate the criminal records of known associates using heuristics, memory, and judgment. For example, when investigating a burglar, a detective may be interested in the drug-related activities of associated individuals. That is, if the target is associated with two different individuals because they were interviewed at a bar fight, the investigator would tend to be more interested in the one with a history of selling drugs. In contrast, previous criminal-association measures simplify records into a network of criminals (nodes) connected by weighted edges characterized by a single measure: association strength.

Suggestions from the analysis module are considered by the analyst as the link chart is formed. The more advanced applications of our importance-flooding algorithm (described in more detail below) recomputed importance based on these accepted/rejected decisions. In addition, while looking at the detailed records of an individual, a detective may discover additional relationships. For example, if a report mentions the name of a sibling, the investigator can add an additional link into the network as the analysis proceeds. It is difficult to assess the correctness of suggestions because judgments can vary substantially from analyst to analyst. The system is considered to be "better" when an analyst is presented with more interesting suggestions earlier in the process.

Importance-Flooding Computation

The importance-flooding algorithm depicted in the center panel of Figure 2 computes an interestingness score for individuals in the network. The basic intuitions of the algorithm are (a) associates of interesting people become relatively more interesting and (b) both a person's past activity and their involvement in interesting association patterns establish initial importance. The algorithm considers two key network elements in its calculation: (a) association closeness and (b) importance evaluation. The calculation leverages association-closeness measures as suggested by Schroeder et al. (2003). Scalar coefficients are implemented as in S. White and Smyth (2003) to dampen the influence of an incident over longer network paths using a decaying distribution function, and a path-based notion of interestingness similar to the methodology used by Lin and Chalupsky (2003) is used to represent more complex heuristics expressed by a crime analyst. The algorithm proceeds in three steps:

1. Relation weights are assigned to network links.
2. Initial importance values are assigned to network nodes.
3. Importance is passed to nearby nodes generating a final score for each node.

Nodes with the highest score are sequentially evaluated by the analyst. Steps 1, 2, and 3 can be repeated after each analyst decision to improve accuracy. Importance flooding employs six components:

1. A set of nodes
2. A set of associations where each association connects two nodes and is described by a set of properties
3. A set of rule-based weights consisting of one link weight for each unique connected pair of nodes
4. Initial importance rules
5. A decaying distribution function
6. A set of starting nodes.

Relation weights. An appropriate link-weighting scheme for analysis of a criminal-activity network can include a number of factors. Building on previous work, our system establishes association strength based on incident records. The association properties we considered include crime type, from-role (the role of the first of the two nodes in the association), to-role (the role of the second node in the association), and crime date. These properties were selected so that we could use a close approximation of the association-strength formula presented in Schroeder et al. (2003). We did not employ shared addresses or phone numbers, even though our methodology allows for this and such information may be useful in our computations. Including these items would increase the complexity of the shared schemas and passing around such data between agencies would introduce additional confidentiality concerns.

Relation weights ranging from 0 to 1 are assigned to each pair of connected nodes in the network. Relation weights are assigned as a function of the number and properties of those associations. We use relatively simple heuristics in the experiments presented here. For example, we assign a strong weight to a pair of individuals when they are both recorded as arrestees in the same incident, but a lower weight when they are categorized as investigational leads in a single incident. Frequency of association is also considered. As suggested by Schroeder et al. (2003), when a pair of individuals appears together in four or more police incidents a maximal relation weight of 1 is assigned regardless of crime role or incident type. When less than four incidents connect two individuals, we multiply the strongest association weight by 3/5, the second strongest by 1/5, and the third strongest by 1/5, and sum the products. The 3/5, 1/5, 1/5 distribution is somewhat arbitrary but it is reasonable in light of previous research.

Initial importance. Initial importance values are assigned to nodes using path-based importance heuristics. In our implementation, we use three kinds of importance rules: (a) activity-based group rules, (b) multi–group membership rules, and (c) path rules, as shown in Figure 3. Weights are assigned to rules, nodes are evaluated for group membership based on the rule, and nodes are assigned initial importance scores equal to the sum of the weights of groups to which they belong. Importance values are normalized to fall

DOI: 10.1002/asi

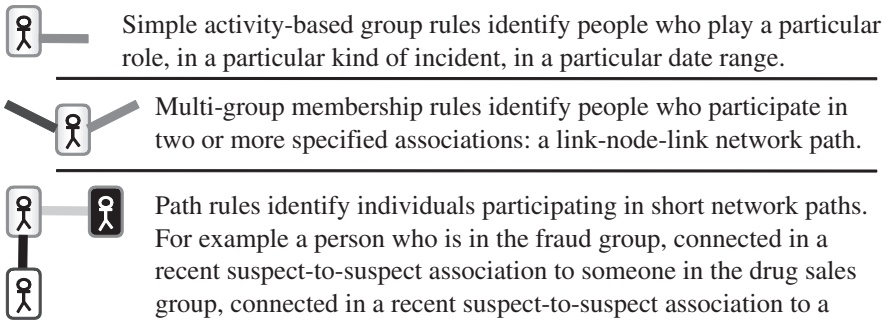

FIG. 3. Three types of initial importance rules.

between 0 and 1 and target nodes are always assigned a score of 1. This model for initial importance was developed in conversation with crime analysts and considers what kinds of information are likely to be shareable in a cross-jurisdictional setting. The path-based heuristics allow analysts to express some of the complex evaluation models used in evaluating case reports. This notion of path-based importance implements a path-based notion of interestingness similar to the results reported in Lin and Chalupsky (2003).

Importance passing. Importance flooding implements the basic notion that a node is more interesting when it is connected to other interesting nodes. Using nodes, links, and initial importance weights as described in the Relation Weights and Initial Importance sections above, importance can be passed through several links, so that if A is connected to B is connected to C, some of A's importance passes to B and a smaller portion passes along to C. As in S. White and Smyth (2003) we use a scalar coefficient to dampen importance passing over longer network paths through a decaying distribution function. If a given person has several identified associates with relatively similar incident histories, the algorithm should guide the analyst towards transitively connected but more interesting individuals. To put it another way, if Bob is associated with Fred, Joe, and Steve, who are relatively unknown, but Steve is associated with John who has a very interesting incident history, the algorithm should suggest Steve before Fred or Joe.

Importance flooding assigns an importance score to nodes as shown in the pseudocode formulation presented in Figure 4. Each node *N1* of *N* has a unique "ID," an initial score "INIT," a previous score "PREV," and an accumulated amount of importance added in this iteration "ADD." The algorithm includes a main loop and a recursive

```
Main Process:
Initialize all nodes N1 in N: N1.PREV= 0, N1.ADD = 0
For each iteration
  For each node N1 in N // Call recursive path tracing
    PassAmt = N1.PREV + N1.INIT
    PathList = N1.ID, PathLen = 1
    pathTrace (PassAmount, PathList, PathLen)
  For each node N1 in N // Normalize and re-initialize
    N1.PREV = (N1.PREV + N1.INIT + N1.ADD) / MAXVAL
    N1.ADD = 0
  // reinforce the importance investigational targets
  For each node T1 in the TargetNode List: T1.PREV = 1

Recursive Path Tracing:
pathTrace (PassAmount, PathList, PathLen)
  PassingNode = The last node included in PathList
  NumOfAssoc = The # of nodes associated with PassingNode
  For each node Na associated with PassingNode
  if Na is not already included in this PathList
    RELWGT = the relation weight for the pair [PassingNode,Na]
    DECAYRATE = scalar coefficient value corresponding to PathLen
    PASSONAMT = PassAmt * RELWGT * DECAYRATE * (1 / NumOfAssoc)
    Na.ADD = Na.ADD + PASSONAMT
      if PathLen < DDD // traverse paths to length DDD
    pathTrace (PASSONAMT, PathList + Na.ID, PathLen + 1)
```

FIG. 4. Pseudocode for the importance-flooding algorithm. The iterative main process calls the recursive pathTrace process, which passes importance from a node to its neighbors over several associational hops as determined by the dampening scalar coefficient.

DOI: 10.1002/asi

path-tracing method. A maximum node importance score of Init + Prev + Add "MAXVAL" is computed to normalize the values after each iteration. A decaying distribution depth "DDD" is used by the computation and is set equal to the number of terms in the scalar coefficient (e.g., if the scalar coefficient is [.5, .25], DDD is 2). While a variety of termination conditions can be used in iterative computations, we simply use a fixed number of iterations. Future work may identify other useful termination approaches. Iteration reduces computation costs. A long scalar coefficient might accomplish similar results but would require much more computation.

Node selection and feedback. Beginning with the investigational targets, the most highly rated, directly connected individual from the pool is suggested as a candidate for inclusion in the link chart. Investigational target nodes are placed into a list of visited nodes and into a priority queue with a priority value of 2. Nodes are sequentially removed from the queue in descending priority-value order. Removed nodes are added to a list of selected nodes and the algorithm scans direct associates. If an associated node is not already in the visited node list, it is added to the priority queue based on its importance score, which can range from 0 to 1. Intuitively, the algorithm asks, "Of all the nodes attached to any of the suggested nodes, which has the highest importance score?"

The network and computation can be adjusted along the way in two treatments we will call "learning" and "discovered-link modes." In the learning mode, when the analyst adds a suggested node to the link chart the initial importance score of the node is increased to 1 (the maximum), and any nodes within two associational hops of the targets are added to the source network. The initial scores of rejected nodes are reduced to 0 but not removed from the network because a person may seem unimportant in and of themselves but still act as a gateway to other important nodes that have not yet been evaluated by the analyst. In the discovered-link scenario, analysts can add additional associations not found in the original data. For example, if a person is added to the link chart and the text of an incident report lists their sibling, that sibling and the strong relationship can be added to the base network to enhance future computations.

Testbed

The data used in our experiments were drawn from incidents recorded by the Tucson Police Department (TPD) and the Pima County Sheriff's Department. The source data includes records from 5.2 million incidents involving 2.2 million people, and had already been converted into a common schema (COPLINK). Associational links were noted whenever two people were listed together in an incident. Based on practitioner suggestions, individuals were matched on first name, last name, and date of birth. Some correct matches were missed due to data-entry errors or intentional deception as they would likely be in any real application of our methodology. To approximate the search space considered by the analyst, we include only people within 2 associational hops of the targets. Investigators tell us they are generally not interested past that limit. We ignored incidents recorded after the chart was drawn.

Manually created link charts from two investigations were considered: Fraud/Meth and Arrow, with 110 people in each. The Fraud/Meth chart (shown in Figure 1) depicts key people involved in fraud and methamphetamine trafficking in the Tucson area, while the Arrow chart focuses on a single criminal investigation. These charts were prepared for the TPD Fraud Unit by a crime analyst who spent several weeks on each chart. The resulting association network included 98/110 people from the Fraud/Meth chart and 100/110 from the Arrow chart. The base network extracted for the Fraud/Meth evaluation (all links between all nodes connected within two associational hops of the targets) included 4,877 nodes and 38,781 reported associations. The Arrow network included 6,025 nodes and 33,574 links. Because one analysis variation employs sensitive query-specific data, we asked the analyst if any additional relationships were discovered and considered that were not included in the database during their investigation. A number of family relationships did figure prominently in the analysis although they were not reflected in the automatically extracted association network. These associations were added and used in the discovered-link scenario described below.

To address the confidentiality, formatting, and administrative concerns that inevitably arise when combining data in a cross-jurisdictional environment, we use a carefully chosen set of attributes: binary relations between two people with a crime type, a date, and the role of each person in the incident. These are items suggested in previous studies for law enforcement (Schroeder et al., 2003) and can be extracted from most records management or law enforcement data warehouse systems. Narratives, addresses, phone numbers, and many other details describing the incident are omitted because they often contain private or sensitive information. Our goal is to demonstrate a useful network-based analysis to identify potential persons of interest from a minimal data set. Investigators would likely follow up by querying existing systems or contacting local agencies.

Experimental Design

To explore the methodology, we implemented eight different ways of ranking nodes with a series of research questions in mind. Table 1 lists our specific research questions, the ranking methods we used (treatments), and hypotheses we tested. Our hypothesis tests compare treatment accuracy using the manually created link charts as a gold standard. Treatments that suggest the "correct" individuals (those selected by the human analyst) earlier in the list were considered to be better. For comparison we use measuring function A, which operates on a ranking method (treatment) over a given size range. As each node is added to a network, we divide the total number of nodes suggested by the number of correct nodes. This ratio represents the number of nodes an analyst would have to

DOI: 10.1002/asi

TABLE 1. Research questions, treatments, and tests.

Research questions Treatments and hypotheses or tests
How well does association-closeness-based analysis support link-chart creation? Does importance flooding improve on closeness-only analysis?
Treatments:
Breadth-first search (BFS) provides a baseline for comparison. Start with the target(s) and choose direct associates, then choose indirect associates. BFS emulates the process an analyst would use when conducting the investigation without analytic support.
Closest associate (CA) applies an adapted shortest-path algorithm. New individuals are suggested in order of association closeness to someone already included in the network.
Importance flooding uses the algorithm described in the Importance-Flooding Computation section in the text to rank nodes. Rankings are not adjusted to respond to analyst selections or discovered links.
Hypotheses:
H1: Importance flooding and closest associate are more accurate than breadth-first search.
H2: Importance flooding is more accurate than closest associate.
Do path-based heuristics add to the accuracy of predictions?
Treatments:
Path heuristics with no flooding employs path-based heuristics to rank importance but does not flood importance to nearby nodes.
Node-only importance flooding employs only simple node-based initial importance rules; no path-based rules are used. Importance is iteratively passed to nearby nodes.
Hypotheses:
H3: Importance flooding is more accurate than node-only importance flooding.
H4: Importance flooding is more accurate than path heuristics with no flooding.
For comparison, what is the best accuracy we can hope to achieve?
Treatment:
Perfect flooding is designed to establish a theoretical upper bound of the importance-flooding algorithm's effectiveness. "Correct" individuals are given initial importance scores of 1 and all others are given 0.
Tests: Comparative accuracy is calculated and reported.
Can we improve results by repeatedly adjusting based on the analyst's selections?
Treatment:
Learning-based flooding adapts its results to user judgments. If a suggested individual is included in the manual chart, the individual's importance is set to 1 (the highest possible value), otherwise the initial importance is reduced to 0. After each correct suggestion, the network is expanded to include individuals found within two associational hops of the correct node. Importance calculations are rerun to reflect these new inputs before additional suggestions are made.
Tests: Comparative accuracy is calculated and reported.
Does the use of links discovered in the course of an investigation improve the result?
Treatment:
Discovered-link results are generated using the learning-based flooding approach with some additional data. Query-specific data such as family relationships, which are missing in the underlying database but available in the text of the incident narrative, are added to the network after one of the people involved has been suggested. When the link identifies a new person who was included in the original chart, that person is immediately suggested and accepted. This approach adds a few new individuals into the network who were not identified in the other treatments.
Tests: Comparative accuracy is calculated and reported.
How sensitive is importance flooding to variation in the computational parameters? How sensitive is importance flooding to variation in the representation of user heuristics?
Tests: Comparative accuracy is calculated and reported.

evaluate for each correct node encountered. A smaller number is better since this means the analyst would have spent less time on uninteresting nodes. Our measure *A* is the average of the ratio over a range. For example, consider *A* (importance flooding) at 250 = average ratio of selected nodes to correct nodes, selected by the importance-flooding algorithm, when the number of selected nodes is 1, 2, 3, . . . 250. Because this measure is undefined (divides by 0) until at least one correct node has been suggested we considered the original target node a correct suggestion.

The first few tests we ran, and all of our hypothesis tests, are intended to generally establish the effectiveness of the importance-flooding algorithm and its components. Each of these hypotheses was tested using the measure *A* described at 100, 250, 500, 1000, and 2000 suggested nodes to see if different treatments or components performed

DOI: 10.1002/asi

differently at different stages of the process. The "breadth-first search" treatment emulates the process an analyst would use when conducting the investigation without analytic support. Although our experimental treatments were expected to improve upon a simple breadth-first search, it is important to remember that there is a chance that the breadth-first search will come across the correct individuals early in the process. The importance-flooding treatments all build on closest-associate analysis in that they use the association-closeness measures as the link weights in the network. At the most fundamental level, importance flooding considers the past behavior of individuals to be a primary indicator of interestingness, while the closest-associate approach only considers past behavior when assessing strength of the relationship. This aligns importance flooding much more closely with the decision process described by investigators.

Our experimentation also considers accuracy limits and how our methodology can adapt to additional input. Table 1 includes a description of the learning-based flooding, discovered-link, and perfect-flooding treatments. An investigator is likely to have access to information that is not available in extracted incident records. In our discovered-link tests, we asked for a list of familial relationships discovered during the analysis. There were approximately 40 additional connections amongst people in our data set as well as a few links to new people. We checked this discovered-link list after each suggestion was made. In the perfect-flooding treatment, the algorithm computes results as if it had perfect importance estimates. Even so, we do not expect perfect results because individuals who were not in the records will still not appear in the suggestions. In addition, although the analyst always identified a direct connection in reality, the records only transitively connect to some of the individuals in the chart. Thus, the perfect-flooding treatment establishes an upper bound of computational accuracy.

The heuristic components used to test the importance-flooding approach came from two sources. Previous research guided the development of the very general link-weight heuristics, and case priorities dictated the importance rules. Each association between a pair of individuals was evaluated: Suspect/Suspect Relationships = .99; Suspect/Not Suspect = .5, Not Suspect/Not Suspect = .3. A single association strength was then assigned as follows: 4 or more associations, weight = 1; else, Σ (strongest relation * .6, 2nd * .2 , and 3rd * .2). Initial importance heuristics included group, multi–group membership, and path rules. Several relevant group rules were identified by the analyst: aggravated assault (A), drug sales (S), drug possession (P), and fraud (F). Individuals were automatically assigned to each group based on events found in the electronic records. Membership in any two of the A, S, and F groups added an importance value of 3 to an individual's total initial importance score, and membership in all three groups added 5. Participation in an A-D-F path added 5 and participation in paths A-D, A-F, D-F, or P-F added 3. For example, in cases where the suspect in an assault A was connected in some incident to a suspected drug seller D who was connected to a suspected check washer F, an initial importance value of 5 was added to each of the nodes. Because of a different focus in the Arrow investigation, two additional group rules were employed: Individuals with a history of both drug possession and theft, or drug possession and burglary, were given an additional initial importance value of 3. The heuristic values represent a numerical approximation of the preferences expressed by the crime analyst.

We conducted a series of tests to see if our technique was robust across variations in the computational parameters, starting points, and numeric representation of user-provided heuristics. Given cost and confidentiality issues, expanding evaluation of the technique over many cases is left for future work. Still, we did conduct some sensitivity testing; results are presented below. It is not clear how to best choose an appropriate scalar coefficient for dampening passed importance over longer network paths, although it might be possible to derive an optimal coefficient by computing and comparing results for a large number of cases. The experiments that compare treatments were run with a scalar coefficient [.5, .25] but we also ran computations using two different coefficients ([.5, .5], and [.75, .5]) with the Fraud/Meth data. Next, we wondered how dependent our technique was on the original list of starting nodes. The treatment-comparing tests were conducted using the target individuals identified by the crime analyst. In addition, we ran our analysis for three alternate, nonoverlapping sets of target nodes. Numerical representations of investigation-specific heuristics are somewhat subjective. So, we tried two alternate formulations to see if they substantially changed suggestion accuracy. We also informally explored different numbers of iterations. We conducted our tests using four iterations because one or two iterations did not seem to produce as good a result and no real differences were observed for more iterations. Although a bit informal (no statistical testing was done because the sample size is relatively small), these tests provide evidence of our technique's robustness over minor variations in the computational parameters, and its promise, even though the same user-provided heuristics can be translated into several numeric representations.

Results

This section presents our experimental results. First we report on our initial attempts to assess the usefulness of importance flooding for link-chart creation. This section includes the hypothesis tests conducted using the Fraud/Meth data as described above. Next, we document the discovered-link, learning, and perfect flooding experiments and discuss the application of these techniques to the Arrow investigation chart. Finally, we present our sensitivity analysis.

Hypothesis Testing: The Fraud/Meth Case

To establish the effectiveness of our approach we compared suggestion accuracy for the importance-flooding algorithm to the breadth-first and closest-associate approaches. Further, we separately tested several components of the

DOI: 10.1002/asi

TABLE 2. Hypothesis testing on the fraud/meth data.

Techniques:
- BFS = breadth first (rank by # of hops)
- CA = closest associate
- IMP = importance flooding
- PATH = path heuristics, no flooding
- NO = only node heuristics, flooding

CA and IMP techniques improve on BFS
- H1a: A(IMP) < A(BFS)
  * *Accepted*
- H1b: A(CA) < A(BFS)
  * *Accepted*

Importance flooding outperforms closest associates
- H2: A(IMP) < A(CA)
  * *Accepted*

Importance flooding outperforms node only heuristics
- H3: A(IMP) < A(NO)
  * *Accepted*

Importance flooding outperforms path heuristics with no flooding
- H4: A(IMP) < A(PATH)
  * *Accepted at 500,1000 & 2000 but NOT 100 or 250*

Hypotheses were tested at 100, 250, 500, 1000, and 2000 selected nodes.
* Accepted hypotheses were significant at $p = .01$

importance-flooding approach to shed light on the source of improvement. Our hypotheses and test results are shown in Table 2. We applied measure A (described in Experimental Design above) when 100, 250, 500, 1000, and 2000 individuals had been suggested by each treatment. All the hypotheses were accepted at all levels except for H4 at 100 and 250 (Table 1). Hypothesis 4 suggests that the combination of flooding and path-based heuristics (IMP) will outperform a treatment that employs path-based heuristics but no spreading activation (PATH). For the first 250 suggestions the IMP treatment was not significantly better at the $p = .01$ or $p = .05$ level. One interpretation of the H3 and H4 tests is that both components, flooding and path-based heuristics, added to accuracy but that the impact of path-based heuristics was especially important in the early stages of the analysis.

Discovered Link, Learning, and Perfect Flooding

Having addressed the suggestion accuracy of the importance-flooding approach in our hypothesis tests, we explored the impact of additional computational input. Performance results for the Fraud/Meth chart are shown in Figure 5 and for the Arrow investigation in Figure 6. In both investigations, all of the importance-flooding approaches consistently found more of the correct nodes for any given number of nodes selected than the closest-associate approach. Perfect-flooding results (nearly all correct suggestions) are included to put the results in context and establish the accuracy limit of the approach. The closest-associate method generally outperformed the breadth-first search in the Fraud/Meth case, and for a good portion of the Arrow computation as well. However, the breadth-first search results outpaced the other methods during a part of the Arrow analysis.

More detail from the Fraud/Meth case is presented in Table 3, which gives the ratio of correct suggestions to total suggestions for the discovered-link, importance-flooding, and closest-associate treatments. This roughly indicates how much effort would be spent reviewing the records of important individuals. For the first 250 suggestions, the discovered-link treatment suggested twice as many important individuals than the closest-associate methodology.

Mean average precision (MAP) is a widely used measure for comparing the accuracy of information-retrieval systems (Harman, 1995). MAP is computed as the mean precision scores measured after each relevant item is suggested (Buckley & Voorhees, 2004). This is a good measure for our task because it gives a relatively higher score when a system returns correct items relatively earlier in the list.

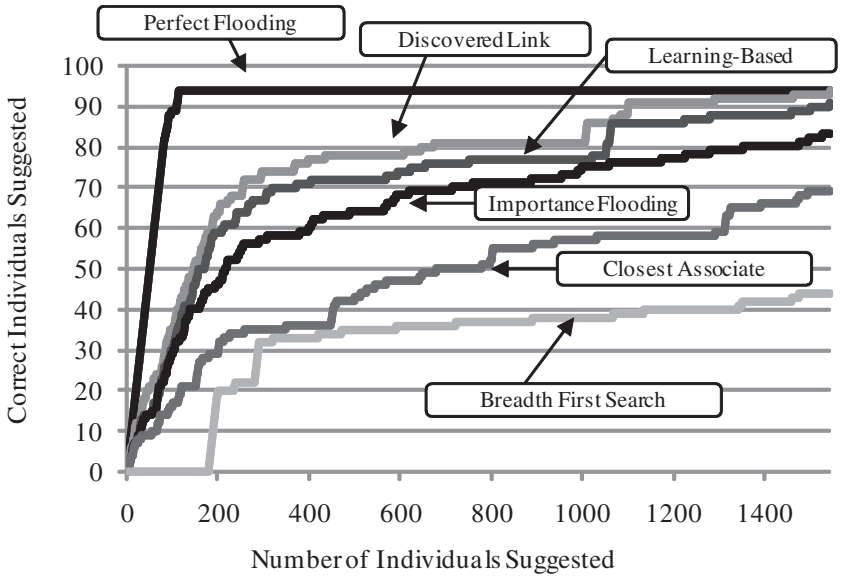

FIG. 5. Fraud/Meth results: Importance flooding (discovered link, learning-based, and importance flooding).

DOI: 10.1002/asi

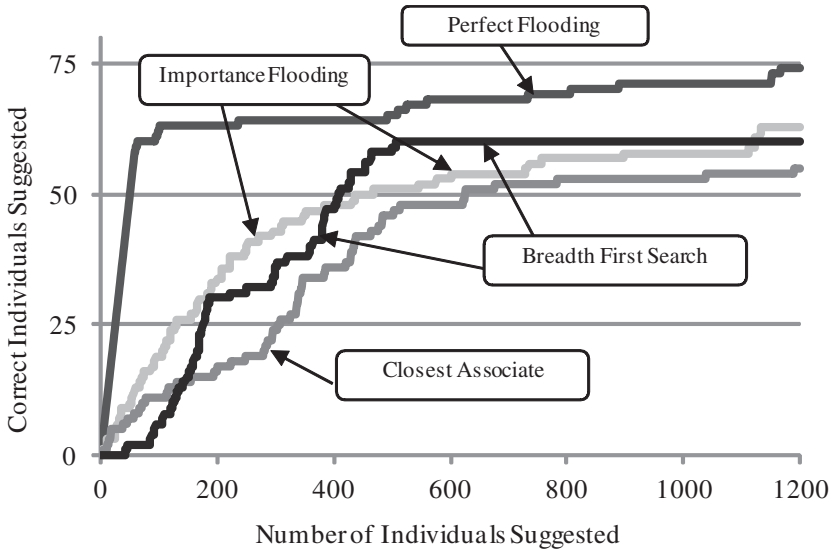

FIG. 6. Arrow investigation results: The importance-flooding approaches (discovered link, learning-based, and importance flooding) had similar results, outperformed the closest-associate approach, and outperformed breadth-first search in finding the first 45 targets. However, the breadth-first search results outpaced the other methods during a part of the analysis.

TABLE 3. Correct suggestions as a percentage of all suggestions.

	Correct suggestions made by various treatments								
	Discovered link			Importance flooding			Closest associate		
People suggested	# correct	% correct	% of network	# correct	% correct	% of network	# correct	% correct	% of network
100	35	35%	36%	29	29%	31%	16	16%	17%
250	68	27%	71%	54	22%	57%	34	14%	36%
500	78	16%	81%	64	13%	68%	54	9%	46%
1,000	81	8%	84%	75	8%	80%	57	6%	61%
2,000	96	5%	100%	85	4%	90%	83	4%	88%

Table 4 shows the MAP values (which are always between 0 and 1) as percentages for all the treatments applied to each of our cases. These results reflect the same pattern seen in Figures 5 and 6.

Sensitivity Analysis

Because the formulation of initial importance heuristics and scalar coefficients is somewhat arbitrary, we explored several variations. In addition, we looked at the possibility that our results were dependent on using the specific starting nodes. We ran the Fraud/Meth case using the learning-based flooding treatment with three different scalar coefficients, four different sets of possible starting nodes, and three variations of the initial importance heuristics as described in the Experimental Design section above. Figure 7 depicts the results, which appear to be relatively consistent despite the various treatments. The three alternate sets of heuristics include the "Original" which is described above; "Alt1," with an added initial importance value of 1 if an individual was a member of the aggravated assault (A), fraud (F), or drug sales (S) groups; and "Alt2," where the heuristics were changed by using a value of 2 where we used 5 in the original computation and 1 where we used 3.

TABLE 4. Mean average precision scores (MAP) for the importance flooding approaches were higher than those for the closest associate and BFS.

Treatment	Fraud/Meth	Arrow
Perfect flooding	99%	83%
Discovered link	33%	15%
Learning-based flooding	26%	15%
Importance flooding	21%	14%
Closest associate	12%	9%
BFS	4%	9%

Discussion

Our results suggest that the importance-flooding computation usefully processed ambiguous networks of interaction

DOI: 10.1002/asi

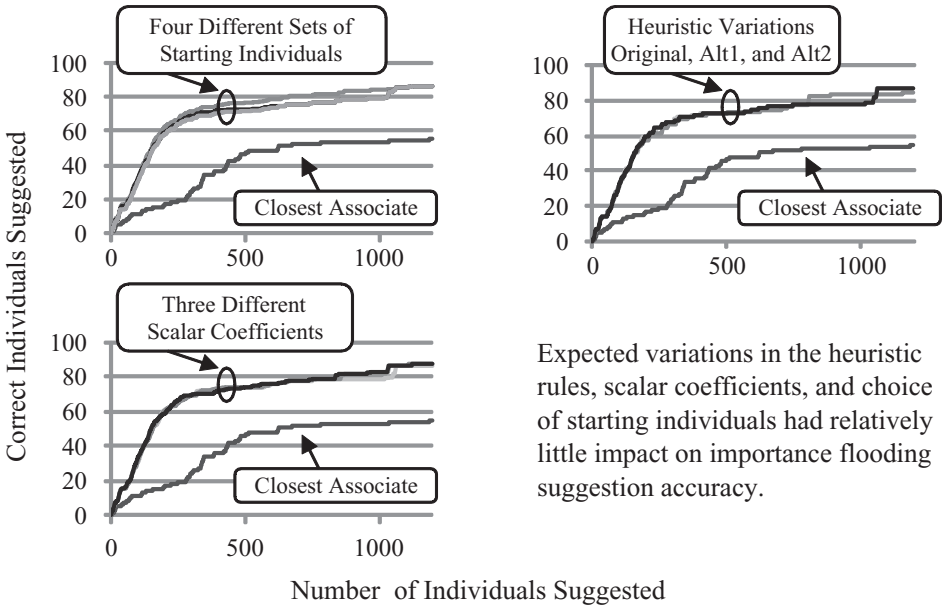

FIG. 7. Fraud/Meth results given parameter variations.

based on a set of user-provided heuristics in support of a realistic analysis task. Our methodology combines several basic notions: (a) association closeness to establish a weighted network of relationships, (b) node-based relevance ranking to account for the recorded actions of individuals, (c) path-based heuristics, which confer importance based on participation in specified patterns of interaction, (d) root-node-based relevance, which focuses computation on a target set of individuals, and (e) iterative spreading activation (flooding), which combines the other elements into a computational model. It is important to note that while our approach is computational, it employs user-provided heuristics for link weights and importance. This is both a strength and a weakness of the approach because, while it allows analysts to explore how different ideas would affect the analysis, poorly conceived ideas might negatively impact the investigational process. Although our testing here is focused on the approach's usefulness in law enforcement, our formulation has implications for other ISI and network-analysis tasks.

This line of work builds on earlier literature by applying the notion of association closeness to the link-chart creation task. We believe that our approach is innovative and promising for the law enforcement domain because it leverages obtainable data, it implements the path-based heuristics used by analysts, it can be adjusted to individual investigations, it is target directed, and it can be meaningfully applied to the important task of link-chart creation. These advances have both theoretical and practical implications. This article expands on our previous publication (Marshall & Chen, 2006) in three important ways: We report on data developed for a second (Arrow) link chart; we include three additional treatments in our experimental results (learning-based flooding, discovered link, and perfect flooding); and we present the results of an initial sensitivity analysis for the importance-flooding algorithm.

Basic Importance-Flooding Accuracy

Our first important finding, based on the results presented in Table 1, Table 2, Figure 5, and Figure 6, speaks to the suggestion accuracy of the importance-flooding methodology, which tended to produce more accurate results as compared to the application of association closeness alone (Hypotheses H1 and H2). We expect that this improved accuracy would be useful in helping analysts identify a higher-quality link chart in a shorter period of time. We note that care should be taken in interpreting these statistical significance tests. Previous studies we reviewed do not propose a methodology for statistically testing differences given a series of decisions in this kind of information-retrieval application.

The statistical analysis reported for Hypothesis H1 seems less compelling in light of the Arrow results. While the closest-associate computation for the Arrow experiment did not always outperform the breadth-first search, we found this result to be less surprising once we considered the case details. The Arrow chart was formed when a particular investigation was already well underway. Instead of beginning with just a few individuals (four in the Fraud/Meth evaluation), the analyst started with 23 of the final 110 individuals. This seems likely to have affected the outcome because many more of the individuals selected for inclusion would be expected to have a recorded and direct association with one or more of the initial targets, thus improving the accuracy of the breadth-first-search approach. The larger number of starting individuals might also have affected the analyst's enthusiasm for exploring longer paths for potentially interesting individuals.

DOI: 10.1002/asi

The acceptance of Hypotheses H3 and H4 suggests that both the flooding and the path heuristics added to the effectiveness of our final result because omitting either technique reduced accuracy. We believe this notion of including heuristics expressed as short paths through a network is important for analyzing graph-based data sets. One could correctly say that a path-based heuristic (e.g., a fraud perpetrator, associated with a drug dealer, who in turn is associated with an enforcer) is really just another characteristic of the individual node. Our approach acknowledges that the "reasoning" appropriate for this task should respect several representational granularities (e.g., a person, a person's previous activities, a person's associations, and a person's participation in generally specified association patterns). This approach moves beyond many previous law enforcement models, which seek to reduce an association graph to a set of weighted links between individuals before applying computational analysis. Systematic inclusion of this kind of heuristic as a precursor to a graph-based computation allows us to better map a crime analyst's notions to the computational model. We speculate that this approach may also be useful in other network-based retrieval applications such as digital library search and user-guided filtering of biomedical relations.

Leveraging Additional Information

We also observe that iteratively adding query-specific information improved suggestion accuracy. Our framework for effectively integrating law enforcement data in support of investigational tasks (first described in Marshall et al., 2004) suggests that because of privacy, safety, data representation, and other issues in this domain, many data are sensitive and cannot be shared across investigations except in ad hoc processes. Thus, we believe that well-designed investigational algorithms need to allow for the inclusion of query-specific data. In the discovered-link scenario, we saw that results improved when we added new links uncovered as the chart was created. Our results indicate that adding these relationships to the computation incrementally improved suggestion accuracy.

User feedback also helped. In the learning-based treatment we adjusted the initial importance scores as individuals were added to the chart. Accepted individuals were boosted to the maximum normalized value of 1, and rejected individuals were reduced to a value of 0. Under this scenario then, a rejected individual could still serve as a conduit through which importance passed to nearby neighbors, but no longer contributes any initial importance of their own to the model. This easily-foreseen adjustment did improve results. We did not systematically test or work hard to optimize for computational performance, but despite the large added computational burden associated with this extra processing, our program (written in Java and running on a standard Windows Pentium 4 desktop PC) was able to produce new suggestions in no more than a few seconds. Repeating the entire iterative computation for each of more than 4,000 suggestions took less than 30 minutes.

Parameter and Heuristic Sensitivity

Our initial sensitivity analysis showed that many expected variations in the operational parameters did not substantially affect results. After some initial testing, we decided to use four iterations in reporting importance-flooding computations. We concluded that running more than one iteration helped, but running more than four did not. We did not perform statistical tests on our exploration of the number of iterations, scalar coefficients, alternate heuristic weights, and different starting groups because we do not believe our data set and test conditions are comprehensive enough to justify such a mathematical treatment. We believe that a convincing and systematic exploration of these parameters would require a much larger set of investigations. Figure 7 demonstrates that our results are not merely a function of a specific choice of parameters and can serve as a starting point for additional investigation.

Law Enforcement Considerations

We tested our methodology using only data that can be realistically generated in the law enforcement domain. The computations in this study use relatively simple relations consisting of a unique identifier for each of a pair of individuals, a coded role identifier for each person, and standard crime-type code for the relation. We did not, for example, attempt to differentiate between drug crimes involving methamphetamines versus drug crimes involving heroin or marijuana, or process textual items such as MO (modus operandi) or physical descriptions. Such details might improve results and could be implemented in our methodology, but they also might be expensive, inconsistent, and subject to additional administrative and privacy restrictions. While many law enforcement organizations would find it possible to share high-level association data (e.g., Bob and Fred were both suspects in an incident of a certain type last June) with certified law enforcement personnel from other jurisdictions, adding more details might make it harder to obtain approval for large-scale sharing efforts. In addition, a number of data-cleaning efforts might have improved our results. For example, we know that the identity-matching rules we used are somewhat crude, but we did not manually adjust our data for even the obvious errors. Wang, Chen, & Atabakhsh (2004) discuss this important issue. We expect that correcting this kind of error in the records would tend to improve our results.

Future Directions

More work can certainly be done to further develop the importance-flooding technique. A larger set of cases would allow further exploration of sensitivity to variations in computational parameters and user-provided heuristics as well as the ability of analysts to effectively express their importance heuristics. We would like to study test cases more deeply to address several practical questions: Are some of the nodes we "suggest" good ones for analysis but left off the charts for a specific reason? Is the technique useful for creating link charts

DOI: 10.1002/asi

with various purposes? Does inclusion of locations, vehicles, and border crossings enhance analysis? We plan to implement some version of the algorithm in a real-time, real-data criminal-association visualization tool to support this kind of detailed work. The value of the approach may increase as data sets grow larger. In our results, the use of path heuristics with no flooding (technique PATH in Table 2) was not significantly different from the complete treatment (technique IMP) until more than 250 nodes were selected. Thus, while the path-based heuristics seem to contribute most to selection value in smaller applications, flooding adds even more value in a larger context. The effect of incorrect identity matching can also be explored. Techniques for consolidating records that contain different identifiers for an individual (because of deception or incorrect and incomplete data) may also impact analysis results and should be explored.

We also plan to adapt importance flooding for other network knowledge representations. The algorithm is designed to overcome link and identifier ambiguity, leveraging a network's structure and semantics. The technique presented here allows us to test this basic notion in other application domains. For example, we plan to explore the use of this algorithm in selecting interesting subsets of a network of biomedical-pathway relations extracted from the text of journal abstracts, and explore its usefulness in matching educational standards to lesson plans and other curriculum elements.

Acknowledgments

This work was supported in part by the NSF Knowledge Discovery and Dissemination (KDD) #9983304, the Information Technology Research (ITR) program COPLINK Center for Intelligence and Security-Informatics Research: A Crime Data-Mining Approach to Developing Border Safe Research, and the Department of Homeland Security (DHS)/Corporation for National Research Initiatives (CNRI) program Border Safe. We are also grateful to Kathy Martinjak, Tim Petersen, and Chuck Violette from the Tucson Police Department for their input.

References

Buckley, C., & Voorhees, E.M. (2004). Retrieval evaluation with incomplete information. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04) (pp. 25–32), New York, NY: ACM.

Chabrow, E. (2002, January 14th). Tracking the terrorists: Investigative skills and technology are being used to hunt terrorism's supporters. Information Week. Retrieved July 7, 2008, from http://www.informationweek.com/news/software/ showArticle.jhtml?articleID=6500694

Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., & Schroeder, J. (2003). COPLINK: Managing law enforcement data and knowledge. Communications of the ACM, 46(1), 28–34.

Coady, W.F. (1985). Automated link analysis: Artificial-intelligence-based tool for investigators. Police Chief, 52(9), 22–23.

Coffman, T., Greenblatt, S., & Marcus, S. (2004). Graph-based technologies for intelligence analysis. Communications of the ACM, 47(3), 45–47.

Gehrke, J., Ginsparg, P., & Ginsparg, P. (2003). Overview of the 2003 KDD Cup. SIGKDD Explorations Newsletter, 5(2), 149–151.

Greenberg, J. (2001a). Automatic query expansion via lexical-semantic relationships. Journal of the American Society for Information Science and Technology, 52, 402–415.

Greenberg, J. (2001b). Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. Journal of the American Society for Information Science and Technology, 52, 487–498.

Harman, D. (1995). Overview of the Fourth Text REtrieval Conference (TREC-4). In The Fourth Text Retrieval Conference (NIST Special Publication 500-236, pp. 1–24) Retrieved July 7, 2008, from http://trec.nist.gov/pubs/ trec4/t4_proceedings.html

Hilderman, R.J., & Hamilton, H.J. (2001). Evaluation of interestingness measures for ranking discovered knowledge. In D. Cheung, G.J. Williams, & Q. Li (Eds.), Lecture Notes in Computer Science: Vol 2035. Advances in Knowledge Discovery and Data Mining (pp. 247–259). Berlin, Germany: Springer.

I2. (2004). I2 Investigative Analysis Software. Retrieved November 29, 2004, from http://www.i2inc.com/Products/Analysts_Notebook/#

Kaza, S., Hu, D., & Chen, H. (2007). Dynamic social-network analysis of a dark network: Identifying significant facilitators. Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI 2007), 40–46.

Kaza, S., Xu, J., Marshall, B., & Chen, H. (2005). Topological analysis of criminal activity networks in multiple jurisdictions. In Proceedings of the 2005 National Conference on Digital Government Research. Marina del Rey, CA: Digital Government Research Center.

Klerks, P. (2001). The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. Connections, 24(3), 53–65.

Koschade, S. (2006). A social-network analysis of Jemaah Islamiyah: The applications to counterterrorism and intelligence. Studies in Conflict & Terrorism, 29, 589–605.

Krebs, V.E. (2001). Mapping networks of terrorist cells. Connections, 24(3), 43–52.

Lin, S.-d., & Chalupsky, H. (2003). Using unsupervised link-discovery methods to find interesting facts and connections in a bibliography dataset. SIGKDD Explorations Newsletter, 5(2), 173–178.

Marshall, B., & Chen, H. (2006). Using importance flooding to identify interesting networks of criminal activity. In S. Mehrotra, D.D. Zeng, H. Chen, B. Thuraisingham, & F.-Y. Wang (Eds.), Lecture Notes in Computer Science: Vol. 3975. IEEE International Conference on Intelligence and Security Informatics (ISI 2006) (pp. 14–25). Berlin, Germany: Springer.

Marshall, B., Kaza, S., Xu, J., Atabakhsh, H., Petersen, T., Violette, C., et al. (2004). Cross-jurisdictional criminal activity networks to support border and transportation security. Proceedings of the Seventh International IEEE Conference on Intelligent Transportation Systems, 100–105.

Melnik, S., Garcia-Molina, H., & Rahm, E. (2002). Similarity flooding: A versatile graph-matching algorithm and its application to schema matching. Proceedings of the 18th IEEE International Conference on Data Engineering (ICDE '02), 117–128.

Miller, G.A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39–41.

NISO. (2005). Guidelines for the construction, format, and management of monolingual controlled vocabularies (ANSI/NISO Z39.19-2005). Bethesda, MD: NISO Press.

Padmanabhan, B., & Tuzhilin, A. (1999). Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems, 27, 303–318.

Raghu, T.S., Ramesh, R., & Whinston, A.B. (2005). Addressing the homeland security problem: A collaborative decision-making framework. Journal of the American Society for Information Science and Technology, 56, 310–324.

Sahar, S. (2002). On incorporating subjective interestingness into the mining process. In Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM), 681–684.

Schroeder, J., Xu, J., & Chen, H. (2003). CrimeLink Explorer: Using domain knowledge to facilitate automated crime association analysis. In H. Chen, R. Miranda, D.D. Zeng, C. Demchak, J. Schroeder, & T. Madhusudan

DOI: 10.1002/asi

(Eds.), Lecture Notes in Computer Science: Vol. 2665: NSF/NIJ Symposium on Intelligence and Security Informatics (ISI 2003) (pp. 168–180). Berlin, Germany: Springer.

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Data and Knowledge Engineering, 8, 970–974.

Sparrow, M.K. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. Social Networks, 13, 251–274.

Wang, G., Chen, H., & Atabakhsh, H. (2004). Automatically detecting deceptive criminal identities. Communications of the ACM, 47(3), 70–76.

White, H.D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. Journal of the American Society for Information Science and Technology, 43, 423–434.

White, S., & Smyth, P. (2003). Algorithms for estimating relative importance in networks. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 266–275). New York, NY: ACM.

Xu, J., & Chen, H. (2003). Untangling criminal networks: A case study. In H. Chen, R. Miranda, D.D. Zeng, C. Demchak, J. Schroeder, & T. Madhusudan (Eds.), Lecture Notes in Computer Science, Vol. 2665: NSF/NIJ Symposium on Intelligence and Security Informatics (ISI 2003) (pp. 232–248). Berlin, Germany: Springer.

Xu, J., & Chen, H. (2004). Fighting organized crime: Using shortest-path algorithms to identify associations in criminal networks. Decision Support Systems, 38, 473–487.

Xu, J., & Chen, H. (2005). Criminal-network analysis and visualization. Communications of the ACM, 48(6), 100–107.

DOI: 10.1002/asi