

USING IMPORTANCE FLOODING TO IDENTIFY INTERESTING NETWORKS OF CRIMINAL ACTIVITY

Byron Marshall¹ and Hsinchun Chen²

¹Accounting, Finance, and Information Management Department, Oregon State University,
Corvallis, OR 97331, USA

byron.marshall@bus.oregonstate.edu

²Department of Management Information Systems, The University of Arizona,
Tucson, AZ 85721, USA

hchen@eller.arizona.edu

Abstract. In spite of policy concerns and high costs, the law enforcement community is investing heavily in data sharing initiatives. Cross-jurisdictional criminal justice information (e.g., open warrants and convictions) is important, but different data sets are needed for investigational activities where requirements are not as clear and policy concerns abound. The community needs sharing models that employ obtainable data sets and support real-world investigational tasks. This work presents a methodology for sharing and analyzing investigation-relevant data. Our importance flooding application extracts interesting networks of relationships from large law enforcement data sets using user-controlled investigation heuristics and spreading activation. Our technique implements path-based interestingness rules to help identify promising associations to support creation of investigational link charts. In our experiments, the importance flooding approach outperformed relationship-weight-only models in matching expert-selected associations. This methodology is potentially useful for large cross-jurisdictional data sets and investigations.

1 Introduction

Events in the last several years have brought new attention to the need for cross-jurisdictional data sharing to support investigations. A number of technology-related initiatives have been undertaken. For example, the FBI sunk \$170 million into a “Virtual Case File” system which was, unfortunately, considered dead on arrival [1]. It will likely be scrapped although lessons learned will benefit future systems. This high profile system failure highlights the difficulty of sharing investigational data across localities. It is even more difficult when multiple agencies are involved, as when local police departments have data of value to national or regional agencies. Computer-supported investigational models are needed to guide the development of policies, protocols, and procedures intended to increase the flow of useful information.

An effective cross-jurisdictional investigation model needs to support real analysis tasks and use data sets that can be realistically collected and shared. In previous work

with the BorderSafe consortium, we developed a model for organizing local data into a network of annotated relationships between people, vehicles, and locations [2]. The proposed methodology considers administrative, policy, and security restrictions aiming to identify a useful data representation that can be collected from existing data sets in spite of administrative and technical challenges. In this paper we explore an importance flooding approach intended to extract interesting CANs (criminal activity networks) from large collections of law enforcement data. Useful analysis models are crucial for the community because without knowing how shared data can be effectively employed, costly resources will likely be wasted in expensive but un-workable integration efforts.

Network-based techniques are commonly used in real-world investigational processes. Criminals who work together in a pattern of criminal activity can be charged with conspiracy and taken off the street for a longer period of time. While many traditional data mining techniques produce un-explainable results, criminal association networks are understandable and actionable. Many networks of associations are “drawn” only in the minds of the investigators, but visual network depictions called link charts are commonly used in important cases. Link charts combine multiple events to depict a focused set of criminal activity. Selected associations may be focused on particular crime types, localities, or target individuals. Link charts are used to focus investigations, communicate within law enforcement agencies, and present data in court. Link chart creation is a manual, expensive, but valuable investigational technique.

An analysis support technique needs to be adaptable because investigational resources are limited and investigational assignments are distributed. Investigators come to a case with specific concerns and relevant experience. Because criminal records are incomplete [3] and missing or ambiguous data such as family relationships are important, rules of thumb (heuristics) need to play a role in analysis if the results are to be accepted by the investigational community. For example, a fraud investigation unit may be only incidentally concerned with drug trafficking. When a crime analyst makes a link chart manually, they look up individual cases, make a judgment as to the importance of particular bits of information, and add information that is not recorded in the regular police records. These investigational parameters change over time. For instance, if the fraud unit realizes that many fraud cases are related to methamphetamine trafficking, they might seek to re-analyze data with an emphasis on this important correlation. Policy concerns also impact analysis. For example, because investigators need to respect individual privacy, law enforcement prefers to focus on individual target(s) rather than “fishing” for patterns in public records.

In any case, one key function of the investigation process is the generation of useful leads. Within this broader context, this paper studies a methodology for increasing the efficiency of link chart creation to (1) save time and money, (2) allow the technique to be used in more investigations, and (3) employ large quantities of available data. Such a model can be used to support investigations and to guide the implementation of data sharing systems. Our research focus can be summarized in a single research question: *How can we effectively identify interesting sub networks useful for link chart creation from associations found in a large collection of criminal incidents employing domain knowledge to generate useful investigational leads and support criminal conspiracy investigations?*

2 Literature Review

Network analysis has a long history in criminal investigation [4-6]. In [3], Sparrow highlights the importance of social network analysis techniques in this important domain, identifying a wide variety of network structure measures and logically connecting those measures with investigational implications. For example, 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?*” or “*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 practitioners.

Some of the analysis techniques anticipated by Sparrow have been explored in more recent work. [6] categorized criminal network analysis tools into three generations. First generation tools take a manual approach allowing investigators to depict criminal activity as a network of associations. Second generation systems include Netmap, Analyst’s Notebook, Watson, and the COPLINK Visualizer [7-9]. 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 would possess advanced analytical capabilities. This class of tool has yet to be widely deployed but techniques and methodologies have been explored in the research literature. [5] 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 [10], and the topological characteristics of cross-jurisdictional criminal networks are studied in [11].

Shortest path measures have received particular attention. One important consideration in an investigation is the identification of the closest associates of target individuals. A variation of this analysis tries to identify the shortest path between two target individuals. These ideas, closest associates and shortest path, are clearly relevant in link chart analysis. CrimeLink Explorer employed relation strength heuristics to support shortest-path analysis [12]. Based on conversations with domain experts, they weighted associations by: (1) crime-type and person-role, (2) shared addresses or phones, and (3) incident co-occurrence. An algorithm for shortest path analysis for criminal networks was implemented and tested in [13]. 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. [13] addresses this using a carefully crafted computational strategy.

Building on this research, we want to help identify “interesting” subsets of large criminal activity networks. The interestingness (or importance) issue is a well recognized problem in the association rule mining field. Interestingness measures seek to assign a ranking to discovered associations based on some interestingness calculation methodology [14]. The various measures of interestingness can be classified into two categories: objective measures and subjective measures [15]. Objective measures are generally statistical and include confidence and support. Subjective interestingness measures, on the other hand, can be classified into two groups: actionable and unexpected. [16] notes that beliefs are important in identifying interesting associations.

Results can be filtered by encoding user beliefs (e.g., expected or potentially actionable relationship or patterns) using some “grammar” and comparing extracted relationships to that grammar [17, 18]. A way to incorporate beliefs is important for automatic interestingness analysis.

Notions of interestingness have received special attention in the context of data that can be represented as a network. Some researchers emphasize that interestingness is relative. For example, a “root set of nodes” within a larger network are used to enhance relevance searching in [19]. They describe a general class of algorithms that use explicit definitions of relative importance. The two main intuitions behind the approach are that 1) two nodes are related according to the paths that connect them, and 2) the longer a path is, the less importance is conferred along that path. Using a scalar coefficient, White and Smyth pass smaller amounts of importance as the distance between a pair of nodes increases. They note several ways of choosing non-overlapping paths between node pairs. These notions of relative importance align well with the cognitive model described by investigators we have talked with. Investigations begin with some target suspect(s) and look for close associates to identify leads.

In [20] novel network paths (not just nodes or links) are identified to reveal interesting information. This was a novel way of analyzing the HEP-Th bibliography data set from the Open Task of the 2003 KDD Cup [21]. Bibliographic citation data 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 co-authors. 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 more 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 is a short path through the criminal activity network.

3 Creating Link Charts by Filtering CANs

Previous work has shown that criminal records can be usefully depicted in a link chart but more advanced methodologies such as criminal network analysis and shortest path evaluation have not been used to directly address the important task of link chart creation. The association rule mining literature suggests several approaches intended to identify interesting items in networks but previous criminal association computations simplify criminal networks using some single measure of association strength. Our

goal is to combine and adapt criminal network and interestingness techniques to support investigational tasks while allowing for the real-world challenges of this important domain. If effective, we expect such a methodology to be useful in a variety of real-world network evaluation applications. Based on our review of the literature and our conversations with investigators we developed a list of design goals:

- 1 Allow query-specific information to fill in missing data.
- 2 Incorporate domain-appropriate heuristics (or beliefs) to support analysis, encoding these heuristics in a format that can be adjusted at query time for new insights.
- 3 Tolerate missing and ambiguous data. While missing information is expected to hamper analysis, a good methodology for this domain needs to be somewhat tolerant of data limitations.
- 4 Be target focused.

Importantly, these goals are applicable to smaller local investigations but are also relevant to large-scale inter-jurisdictional investigations.

We propose the use of an importance flooding algorithm to identify interesting sub networks within larger CANs to help detectives interactively construct investigational link charts. This represents one phase of a larger process in which police records are organized for sharing as described in [2]. Police records from local jurisdictions are converted into a common schema. Person records are matched to form a network of incident-based associations. Then, with a target list of suspects and sets of link weight rules and importance heuristics, individuals are importance ranked for inclusion in investigation-specific link charts. The basic intuitions of the algorithm are (1) associates of interesting people become relatively more interesting and (2) 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 (1) association closeness and (2) importance evaluation. The calculation leverages association closeness measures as suggested by [12], scalar coefficients as in [19], and leverages a path-based notion of interestingness reminiscent of the methodology used in [20]. The algorithm proceeds in four basic steps:

1. 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.
4. A network subset is selected starting with target nodes and best first search.

Our algorithm employs 6 components: a set of nodes, a set of associations such that each association connects two of the nodes and is described by a set of properties, a set of rule-based relation weights consisting of a single link weight for each unique pair of nodes connected in the associations, initial importance rules, a decaying distribution function, and a set of starting nodes.

In this paper, we test our approach by comparing the output of an importance flooding computation with two link charts which had previously been created by a crime analyst from the Tucson Police Department (TPD). The nodes in the network we test in this work are individuals found in an integrated TPD/Pima County Sheriff's Department data set. We used only people as nodes although the algorithm could also evaluate location or vehicle entities. 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 [12].

Relation weights ranging from 0 to 1 are assigned to each association found in the records. Relation weights are assigned to node pairs by evaluating the corresponding associations as a function of the number of associations and properties of those associations. We used relatively simple heuristics in the testing presented here. For example, our rules expressed a strong relational weight for a pair of individuals who were both recorded as arrestees in the same incident, but a lower weight for associations where the two individuals were considered investigational leads in the same incident. In addition to these initial link weights, frequency of association was considered. As suggested by [11], 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 values are assigned to nodes using path-based importance heuristics. In our current implementation, we accept three kinds of importance rules: (1) activity-based group rules, (2) multi-group membership rules, and (3) path rules. Figure 1 describes the three types of rules. Weight values are assigned to each rule, each node is evaluated for group membership based on the rule, and a node is assigned an initial importance score equal to the sum of the weights of all groups to which the node belongs. Importance values are normalized to fall between 0 and 1 and target nodes are always assigned a score of 1. The link weight and importance values assigned in our implementation were derived from previous research or developed in conversation with crime analysts and require only information that is likely to be available in a cross-jurisdictional setting.

In these experiments, our decaying distribution function used .5 for any directly connected nodes and .25 for transitively connected nodes and the target nodes are identified by the analyst. Pseudo code for the iterative importance flooding calculation is shown below. Each node N_1 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 path tracing

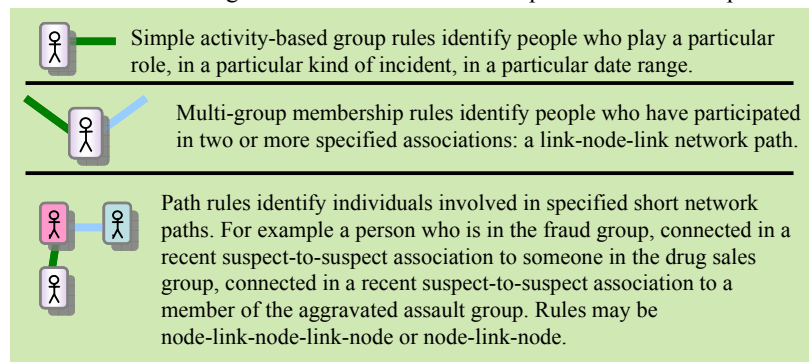


Fig. 1. Three Types of Initial Importance Rules

method. A maximum node importance score of $\text{Init} + \text{Prev} + \text{Add}$ "MAXVAL" is maintained for each iteration as each node and path is processed. This score is used to normalize the values at the end of 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).

```

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 PathList
      RELWGT = the relation weight for the pair [PassingNode,Na]
      DECAYRATE = the decay coefficient for PathLength
      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)

```

Finally, a best first search algorithm uses the resulting importance scores to expand the network from the target nodes to a network of some specified size. The nodes in the starting list of target nodes are placed into a list of visited nodes and into a priority queue with a priority value of 2. Nodes are sequentially popped from the queue until enough nodes have been selected. As each node is popped, the algorithm adds it to a list of selected nodes and then searches for all other nodes associated with that node. If the associated node is not already in the visited node list, it is added to the priority queue with its importance score (which can range from 0 to 1) as its priority value. Intuitively, the algorithm asks: of all the nodes attached to any of the selected nodes, which has the highest importance score? An analyst using the output might well consider which node to add to a link chart next using a similar procedure.

4 Experimentation

To explore the usefulness of our methodology we needed a human-generated link chart and a large criminal activity network, along with heuristics and targets for a par-

ticular investigation. We obtained access to a large link chart prepared for the TPD fraud unit. It depicts key people involved in both methamphetamine trafficking and fraud. The chart includes 110 people and originally took 6 weeks to create.

We drew our network from incidents recorded by the Tucson Police Department and the Pima County Sheriff's Department. The records were converted to a common schema (COPLINK) and associations were created whenever two people were listed in an incident. We recorded "crime type," "from role" (the 1st person's role), "to role" (the 2nd person's role), and "crime date." Using practitioner-suggested guidelines, 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. The combined set includes records from 5.2 million incidents involving 2.2 million people. 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 records added after the chart was drawn. This filtering process resulted in 4,877 individuals for the fraud/meth investigation, including 73 of the 110 "correct" individuals depicted in the manually created link chart.

The heuristic components 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, \sum (strongest relation * .6, 2nd * .2, and 3rd * .2). Initial importance calculations included group, multi-group, and path rules. Several relevant groups were identified by the analyst: Aggravated Assault (A), Drug Sales (S), Drug Possession (P), Fraud (F). Membership in any of these groups added an importance value of 3 to an individual's total initial importance score. Membership in any two groups added 3 more, and membership in all three groups added 5. Participation in an (A)-(D)-(F) 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.

We compared our algorithm's results to the human-drawn link chart, considering how the algorithm might impact the effectiveness of time spent working on the link chart. When an analyst creates a chart, they begin with one or more target individuals, look for associations involving those individuals, and evaluate each potential associate to see if they are important enough to be included in the chart. Reviewing more promising associates first would allow creation of a good chart in less time. In our tests we started with the same information considered by the human analyst and produced an ordered list of individuals such that selecting them in order forms a network. Selection methodologies that listed the "correct" individuals (those selected by the human analyst) earlier in the list were considered to be "better." We compared several methods of ordering the lists, including several variations of importance flooding:

- *Breadth First Search* provides a baseline for comparison. Start with the target(s) and choose direct associates, then choose indirect associates.
- *Closest Associate* is a link-chart application of previously proposed shortest path algorithms. New individuals are added to the network in order of association closeness to someone already included in the network.

- *Importance Flooding* was used to rank all the individuals. New individuals are added to the network by choosing the highest ranked individual associated with any of the people already included in the network.
- *Path Heuristics with No Flooding* employed the path-based heuristics to rank importance but did not flood importance to nearby nodes. This was intended to show that both the initial importance of a node and its structural place in the network impact its chart-worthiness.
- *Node-only Importance Flooding* demonstrated that the path-based heuristics add to the algorithm's effectiveness as a supplement to node-only analysis.

For comparison we used measuring function A which operates for a ranking method (technique) over a size range. As each node is added to a network, we can compute the total number of nodes added divided by the number of "correct" nodes added. This ratio computes the number of nodes an analyst would have to consider for each correct node considered. A smaller number is better in that the analyst would have spent less time on un-interesting nodes. Our measure A is the average of the ratio over a range. For example, consider $A(\textit{importance flooding}) \textit{ at } 250 = \textit{average ratio of selected nodes to "correct" nodes, selected by the importance flooding algorithm, when the number of selected nodes is } 1,2,3 \dots 250$. Our hypotheses are shown in Table 1.

Table 1. Hypotheses

Techniques:	<ul style="list-style-type: none"> • IMP = importance flooding • BFS = breadth first (rank by # of hops) • CA = closest associates • PATH = path heuristics, no flooding • NO = only node heuristics, flooding
All techniques improve on BFS	<ul style="list-style-type: none"> • H1a: $A(\text{IMP}) < A(\text{BFS})$ * <i>Accepted</i> • H1b: $A(\text{CA}) < A(\text{BFS})$ * <i>Accepted</i>
Importance flooding outperforms closest associates	<ul style="list-style-type: none"> • H2: $A(\text{IMP}) < A(\text{CA})$ * <i>Accepted</i>
Importance flooding outperforms path heuristics with no flooding	<ul style="list-style-type: none"> • H3: $A(\text{IMP}) < A(\text{PATH})$ * <i>Accepted at 500,1000 & 2000 but NOT for 100,250</i>
Importance flooding outperforms node only heuristics	<ul style="list-style-type: none"> • H4: $A(\text{IMP}) < A(\text{NO})$ * <i>Accepted</i>
Hypotheses are expected to hold for 100, 250, 500, 1000, and 2000 selected nodes. * <i>Accepted Hypotheses were significant at $p=.01$</i>	

Performance results for the basic methods (breadth first, closest associate, and importance flooding) are reported in Figure 2. The importance flooding approach consistently found more of the correct nodes for any given number of nodes selected. The closest associate method seems to have generally outperformed breadth first search. In addition, based on the acceptance of hypotheses 3 and 4, we observe that both the flooding and the path heuristics added something to the effectiveness of our final result because omitting either part reduced accuracy. When a second link chart was also analyzed, the importance flooding algorithm again outperformed the best first search and closest associate methods. Detailed results are omitted because of space limitations.

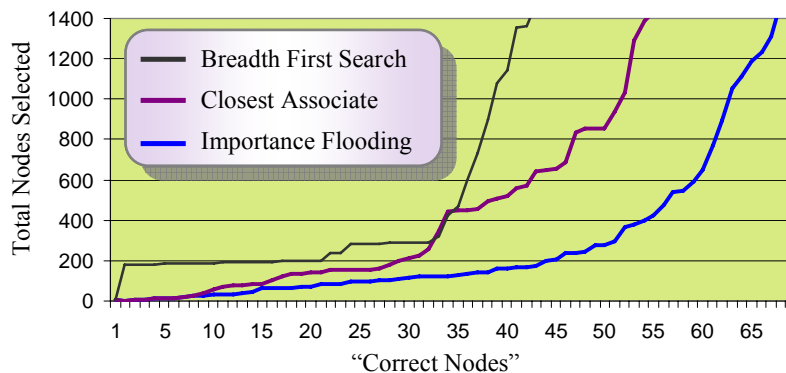


Fig. 2. Comparison of Ranking Methods for the Fraud/Meth Link Chart. The importance flooding algorithm (blue) consistently outperformed other methods.

5 Discussion and Future Directions

Our approach differs from previous work in several ways. (1) It is applied directly to the task of link chart generation. Previous work has hinted at this kind of application but has not experimented with actual charts. (2) We combine structure (closeness-weighted associations) and activity-based importance heuristics (e.g. “people who have been involved in fraud”) in our computation instead of social network measures based on closeness-weighted associations. (3) We encode the users’ importance notions as short network paths. This can be simple grouping (e.g. people who have been suspects in fraud incidents) but we also leverage relational patterns. For example, one of the heuristics we use in our testing process captures the analyst’s input that she was more interested in people who sold drugs and were associated both with people who sold methamphetamines and people who committed fraud. (4) Our approach is target-directed. These advances have both theoretical and practical implications.

We tested our methodology using data that could be realistically generated in the law enforcement domain. The network representation used in our study can be (and was) generated from actual criminal records systems recorded in different records management systems in different jurisdictions. Our methodology does not require analysis of difficult to process items such as MO (modus operandi) or physical descriptions. What’s more, our current representation categorizes crimes using standard crime types which do not differentiate, for instance, between drug crimes involving methamphetamines vs. drug crimes involving heroine or marijuana. Certainly these features can play an important investigational role but extraction of such details might be expensive, inconsistent, and subject to additional administrative and privacy restrictions in a cross-jurisdictional environment. Our results demonstrate analysis value in spite of limited representational detail. With all that being said, additional features could be used by the algorithm simply by changing the initial input rules. We believe the association model we propose (entities connected in labeled associations including roles, types, and dates) is flexible enough to support various investigational tasks, yet

simple enough to be readily sourced from different underlying records management systems. Different analysis implementations could leverage different feature sets when the needed data was relevant and available. But even when association details cannot be shared because of policy, financial, or technical limitations, we believe many organizations would find it possible to share high level association data (e.g. Bob and Fred were both involved in a drug investigation last June) with certified law enforcement personnel from other jurisdictions.

While promising, our results need further validation. Because of restrictions on the sharing of information about old investigations, we only tested on two link charts. Even then, the nodes included in the manually prepared link chart are a “bronze standard” rather than a “gold standard.” It may be that some people “should” have been included but were not because they were missed by the analyst or left off for a variety of reasons. If an individual was in prison or was working with the police as an informant, they may have been omitted from the chart. Thus we have no real objective standard to say that one chart is “correct” while all others are “incorrect.” Instead we would argue that some charts are clearly better than others. Also, sensitivity to variations in computational parameters and user-provided heuristics should be explored.

More work can certainly be done in the law enforcement domain. 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? How much can we improve results by adding query specific data to the importance ranking calculations? Is the technique useful for creating link charts 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 1) 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 to selection value in smaller applications, flooding adds even more value in a larger context.

We plan to test importance flooding in other informal node-link 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.

6 Acknowledgements

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

7 References

1. Schmitt, R.B., New FBI Software May Be Unusable, in Los Angeles Times. 2005: Los Angeles, CA.
2. Marshall, B., et al. Cross-Jurisdictional Criminal Activity Networks to Support Border and Transportation Security. in 7th International IEEE Conference on Intelligent Transportation Systems. 2004. Washington D.C.
3. Sparrow, M.K., The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects. *Social Networks*, 1991. **13**(3): p. 251-274.
4. Coady, W.F., Automated Link Analysis - Artificial Intelligence-Based Tool for Investigators. *Police Chief*, 1985. **52**(9): p. 22-23.
5. Coffman, T., S. Greenblatt, and S. Marcus, Graph-Based Technologies for Intelligence Analysis. *Communications of the ACM*, 2004. **47**(3): p. 45-47.
6. Klerks, P., The Network Paradigm Applied to Criminal Organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. *Connections*, 2001. **24**(3): p. 53-65.
7. Chabrow, E., Tracking The Terrorists: Investigative skills and technology are being used to hunt terrorism's supporters, in *Information Week*. 2002.
8. I2. I2 Investigative Analysis Software. 2004 [cited 2004 November 29]; Available from: http://www.i2inc.com/Products/Analysts_Notebook/#.
9. KCC. COPLINK from Knowledge Computing Corp. 2004 [cited 2004 November 29]; Available from: <http://www.coplink.net/vis1.htm>.
10. Xu, J. and H. Chen. Untangling Criminal Networks: A Case Study. in NSF/NIJ Symp. on Intelligence and Security Informatics (ISI). 2003. Tucson, AZ: Springer.
11. Kaza, S., et al., Topological Analysis of Criminal Activity Networks: Enhancing Transportation Security. *IEEE Transactions on Intelligent Transportation Systems*, forthcoming, 2005.
12. Schroeder, J., J. Xu, and H. Chen. CrimeLink Explorer: Using Domain Knowledge to Facilitate Automated Crime Association Analysis. in *Intelligence and Security Informatics, Proceedings of ISI-2004, Lecture Notes in Computer Science*. 2003: Springer.
13. Xu, J. and H. Chen, Fighting Organized Crime: Using Shortest-Path Algorithms to Identify Associations in Criminal Networks. *Decision Support Systems*, 2004. **38**(3): p. 473-487.
14. Hilderman, R.J. and H.J. Hamilton, Evaluation of Interestingness Measures for Ranking Discovered Knowledge. *Lecture Notes in Computer Science*, 2001. **2035**: p. 247-259.
15. Silberschatz, A. and A. Tuzhilin, What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Transactions on Data and Knowledge Engineering*, 1996. **8**: p. 970-974.
16. Padmanabhan, B. and A. Tuzhilin, Unexpectedness as a Measure of Interestingness in Knowledge Discovery. *Decision Support Systems*, 1999. **27**(3): p. 303-318.
17. Sahar, S. On Incorporating Subjective Interestingness into the Mining Process. in *Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE International Conference on*. 2002.
18. Sahar, S. Interestingness Preprocessing. in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. 2001.
19. White, S. and P. Smyth. Algorithms for Estimating Relative Importance in Networks. in *ACM SIGKDD intern'l conference on knowledge discovery and data mining*. 2003. Washington, D. C.: ACM Press.
20. Lin, S.-d. and H. Chalupsky, Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset. *SIGKDD Explor. Newsl.*, 2003. **5**(2): p. 173-178.
21. Gehrke, J., P. Ginsparg, and P. Ginsparg, Overview of the 2003 KDD Cup. *SIGKDD Explor. Newsl.*, 2003. **5**(2): p. 149-151.