



# Relational Mining for Compliance Risk

A Presentation for the MTC

Eric J. Toder  
Director, Office of Research  
Internal Revenue Service

July 2004

# Outline

- Flow-Through Tax Entities
- Evolution of IRS Data Mining Efforts
- IRS Suite of Tools
  - Currently in Use:
    - Link Analysis
    - Graph Query Tool
  - Plans for future tools and technologies

# Flow-Through Entities Overview

- Partnerships (Form 1065)
- Subchapter S Corporations (Form 1120S)
- Trusts (Form 1041)
- Common characteristics:
  - pay no taxes at the entity level
  - use forms K-1 to report income, credits and deductions to partners, shareholders and beneficiaries
- K-1s link flow-through entities, creating tiering structures
- Use of flow-through entities is suspected in abusive schemes to hide, manipulate, or offset income on taxable returns

# Evolution of IRS Efforts

- IRS transcribes Schedule K-1 documents for the first time for tax year 2000
- Market review and technology assessment proposed using Link Analysis Technology (August 2002)
  - IRS Office of Research awarded a proof of concept contract to MITRE Corp, the IRS Federally Funded Research & Development Center
- Proof of concept for use of link analysis for flow-through entities proven very quickly (May 2003)
- NHQ Research funds relational mining effort at MITRE through November 2004

# Suite of Tools

The effort will build a collection of analytical tools specifically designed to help the IRS explore relationships between taxpayers. Currently, Link Analysis is focused on flow-through relationships created by Partnerships, Trusts and Subchapter-S Corporations.

Tool development is being overseen by NHQ Research Group in collaboration with Operating Divisions Research Staff using the MITRE Corporation as a contractor.

# What is Link Analysis?

- Link analysis is the process of building networks of interconnected objects through relationships in order to expose patterns and trends
- Link Analysis provides a powerful means of performing visual data mining, particularly if you know how to take advantage of layout options, filter assessments and presentation formats
- Link Analysis systems allow you to identify patterns, emerging groups and related connections quickly

# Uses of Link Analysis

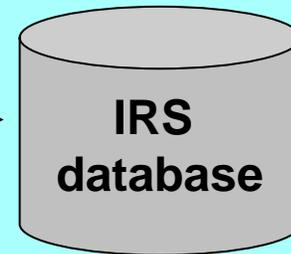
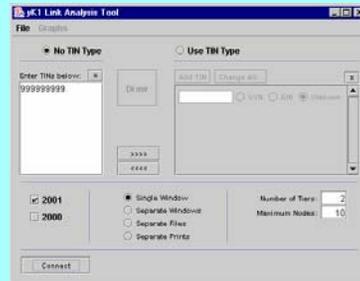
- Provides the ability to identify:
  - Patterns with no business purpose
  - Planned versus incidental relationships
  - Strong relationships and controlling interests
  - Abusive use of tiering such as disappearing money
- Determine holistic view of economic entity
- Need for abstraction and structure analysis

# What is yK1?

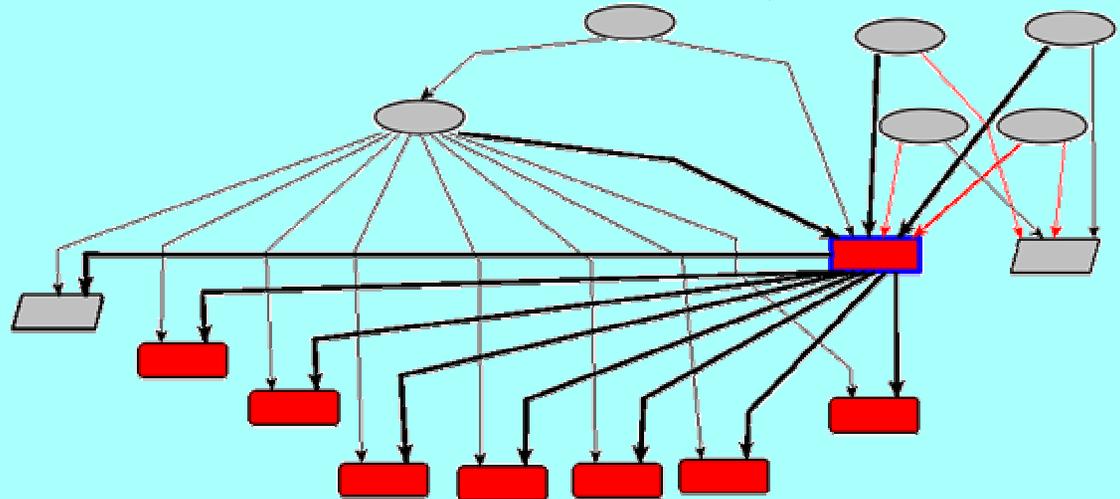
- yK1 is the interactive link analysis tool developed by IRS Research to discover and explore tax entities and their relationships
- Application draws graphs of nodes and links
  - Nodes represent tax returns and other entities
  - Links between the nodes reflect K-1 activity and other connections between entities (for example, parent-subsidiary or husband-wife relationships)
- Nodes and links can be queried to provide more detailed information
- Graphs can be expanded by iteratively pursuing nodes of interest

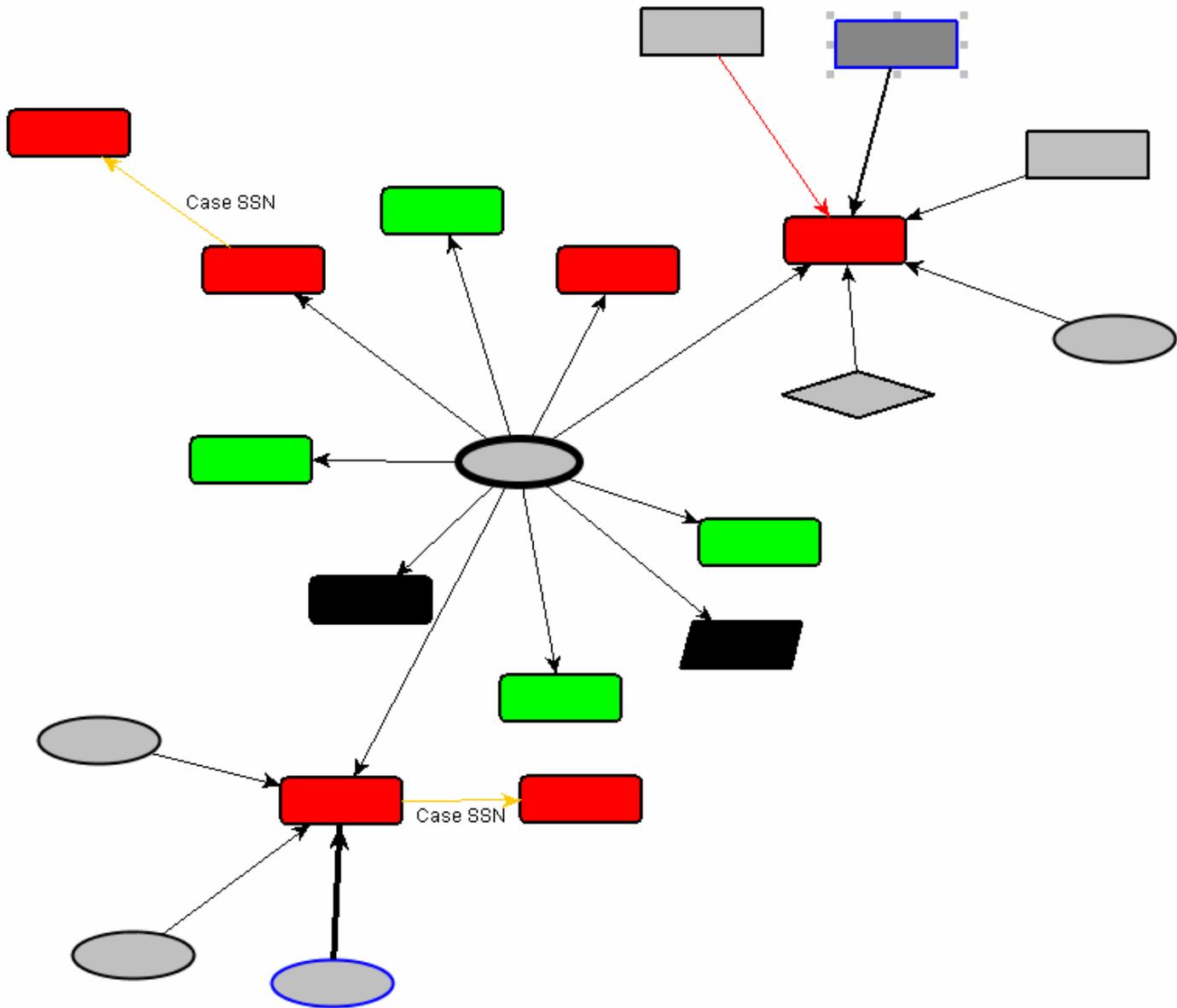
# Pattern Visualization and Investigation, yK1

**Visualization  
Request for  
TIN:**



**Investment  
Structure Pattern  
of TIN and  
Related Entities:**





# Relational Data Mining

- Relational mining
  - Instead of focusing on one return at a time, consider indicators from multiple forms filed by multiple entities
    - Example: Not just the 1040 (individual tax return) and/or the 1041 (trust tax return) and/or the 1065 (partnership tax return) and/or the 1120S (S corporation tax return), but all, especially focusing on key interrelationships
- Sample tasks:
  - Given an example of an Abusive Tax Avoidance Transaction, identify similar examples (allowing the user to refine the pattern interactively)
  - Describe observed patterns of interaction

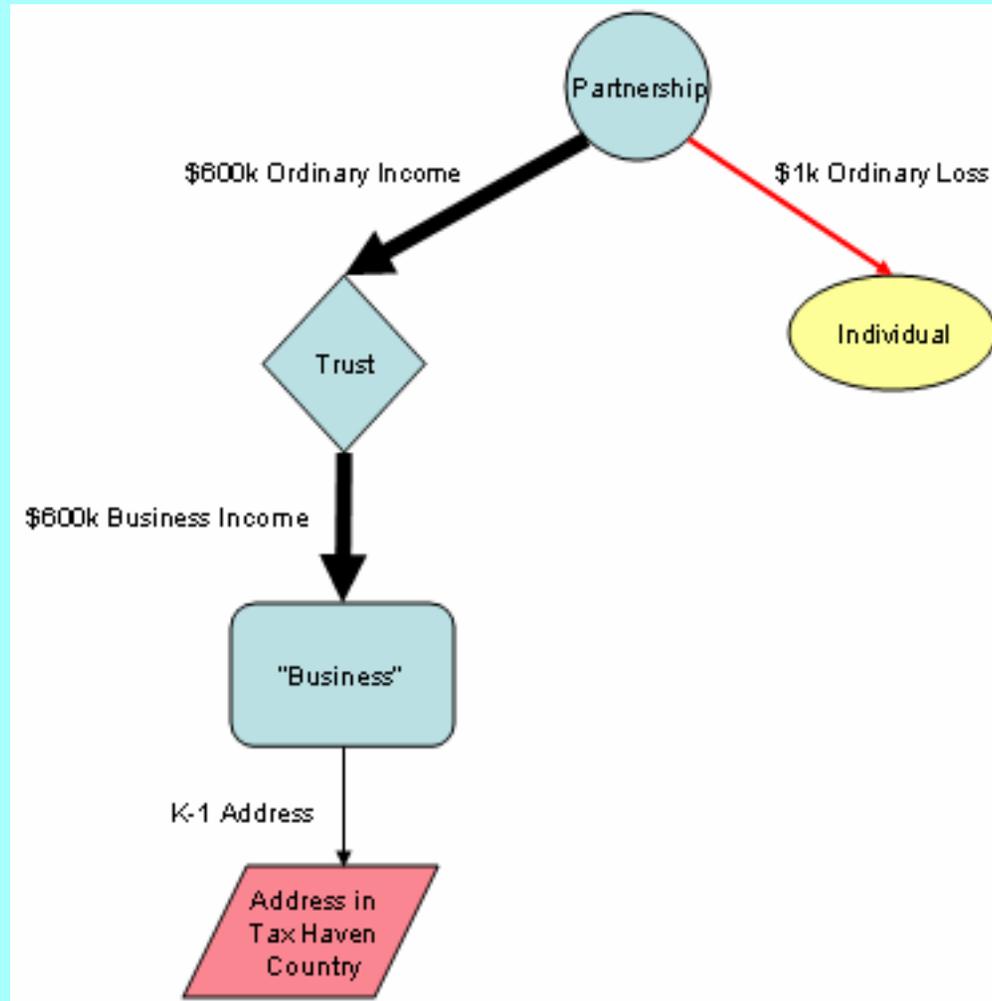
# Reactive and Proactive Analyses

- Reactive Analysis (Bottom Up)
  - Analyst begins with a specific hypothesis.
  - Used when a specific entity (person, location, organization, etc.) is targeted from the outset.
  - The target and all related information are extracted for analysis and paths of connections between the target and other entities are explored.
  - Useful pictures to describe relationships.
- Proactive Analysis (Top Down)
  - No specific hypothesis identified a priori; Goal is to discover previously unknown patterns and trends.
  - Begin by sampling to get overall picture; pursue interesting paths as they arise by re-sampling to narrow the search to investigate emergent hypotheses.

# Use of Link Analysis and Relational Data Mining

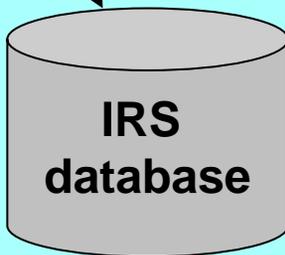
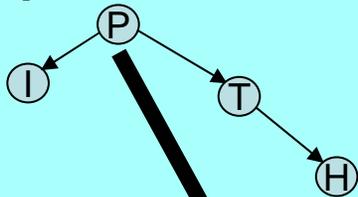
- Link analysis - Reactive
  - Interest in specific person or corporation
  - Focused line of investigation
  - Requires a starting point, result is set of related networks
  - Best to be done interactively
  - Typically done as part of investigative methodology
- Relational data mining -Proactive
  - Interest in a broad-based class of K-1 networks
  - Typically, part of a strategic analysis process
  - Iterative process to refine data and work through anomalies
  - Can be used in classification mode to understand and group data, find outliers
  - Usually done with advanced computer algorithms

# Graph Showing An Individual Connected to a Tax Haven Address

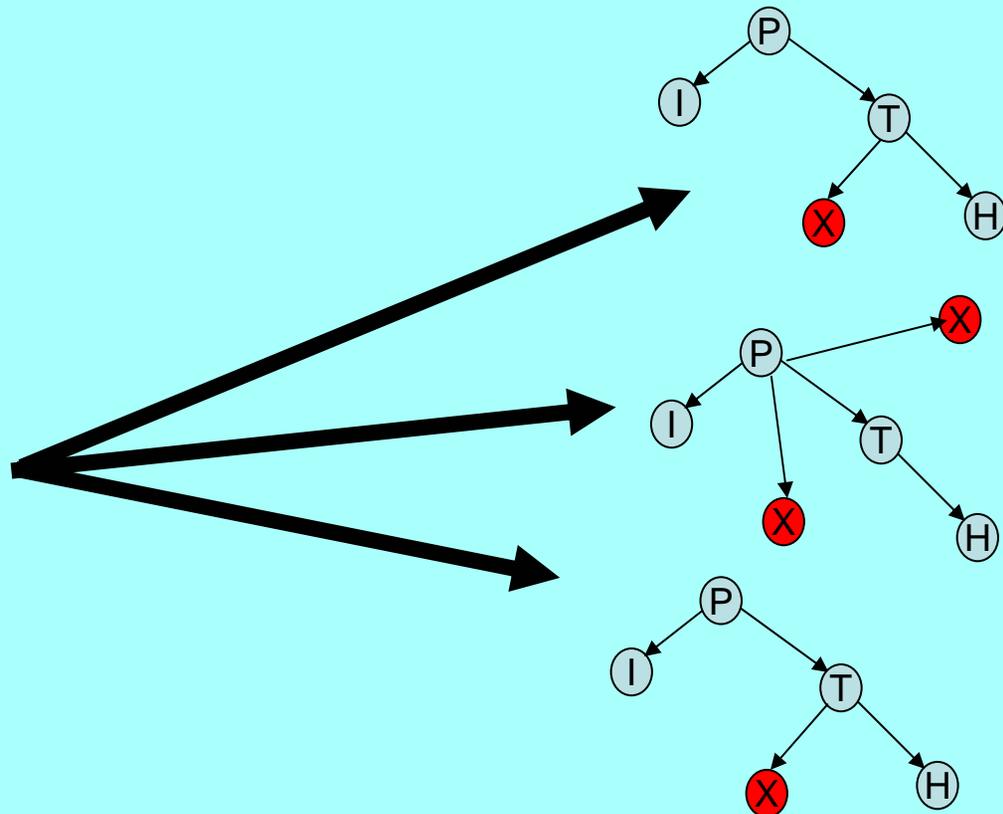


# Graph Query Pattern Matching

Input query pattern:



Patterns that match:



# Status of IRS Relational Mining Efforts

- Development of Specialized Data Structures and Algorithms
  - Requires iterative interviews with IRS Domain Experts
  - Customized model building and refinement
- Visualization tool (yK1) being actively tested by several users
- Graph Query Pattern Matching Tool prototype is being evaluated
- Test cases delivered for classification/audit
- Several diverse domain areas and experts have been polled and recommendations for further development have been submitted by MITRE – Report to be completed
- Several research areas/algorithms to solve these problems are being actively pursued

# Legend

## File

1065 Partnership

1041 Trust

1120S Sub-S

1120 Corporation

Other 1120 1120 Variant (e.g. 1120REIT)

TEGE Form 990 or 5227

1040 Individual

??? Unknown

Group Node Payees, Payers, Subs, etc.

Green ITIN

Black Haven Address

Red IRTF-TPI > \$250,000

Red BRTF-Total Assets > \$250,000,000

Yellow Shelter

Bold Border Initial TIN

Blue Border Form 5500 Filed

Positive Flow

Negative Flow

Parent / Sub Link

Case SSN Link