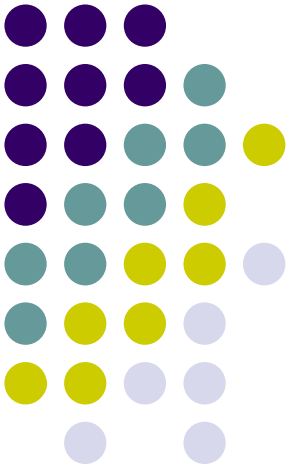
Data Analytics Tools to Support Counter-Terrorism



Hans Chalupsky

Project Leader, Loom KRR&D Group
Information Sciences Institute
University of Southern California
hans@isi.edu

Joint work with:
Jafar Adibi, Aram Galstyan, Shou-de Lin, Eric Melz, Tom Russ, Andre Valente

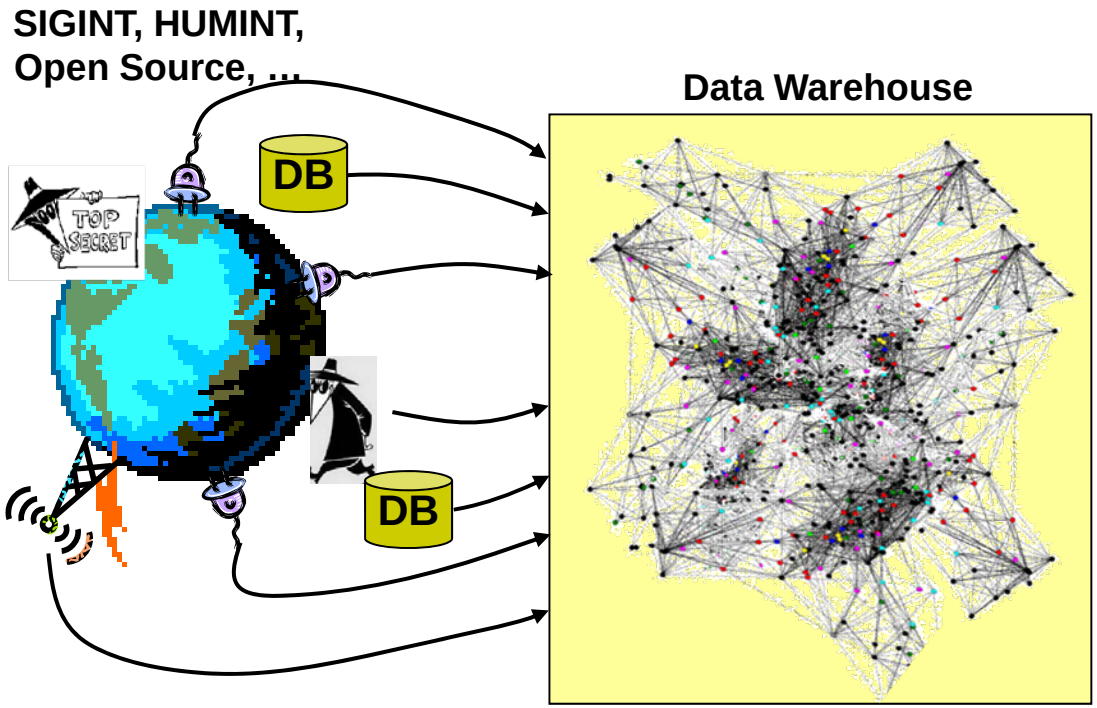
Counter-Terrorism Challenges

- Relatively Rare Events
  - In 2009, 4 incidents in the U.S. as reported by NCTC's incident tracking system
  - But: 1,827 in Iraq, 1,404 in Afghanistan, 8,260 world-wide (2009)
- Low Profile
  - Small, autonomous cells, difficult to detect "micro actors"
  - Madrid train bombing cost less than $30,000
- Long Term
  - Planning of 9/11 attack started in 1995
- Highly Adaptive, Evolutionary
  - "…enemy's proven ability to adapt…"
  - "…likely that we will face a resilient enemy for years to come." State Dept. Country Reports on Terrorism, 2005

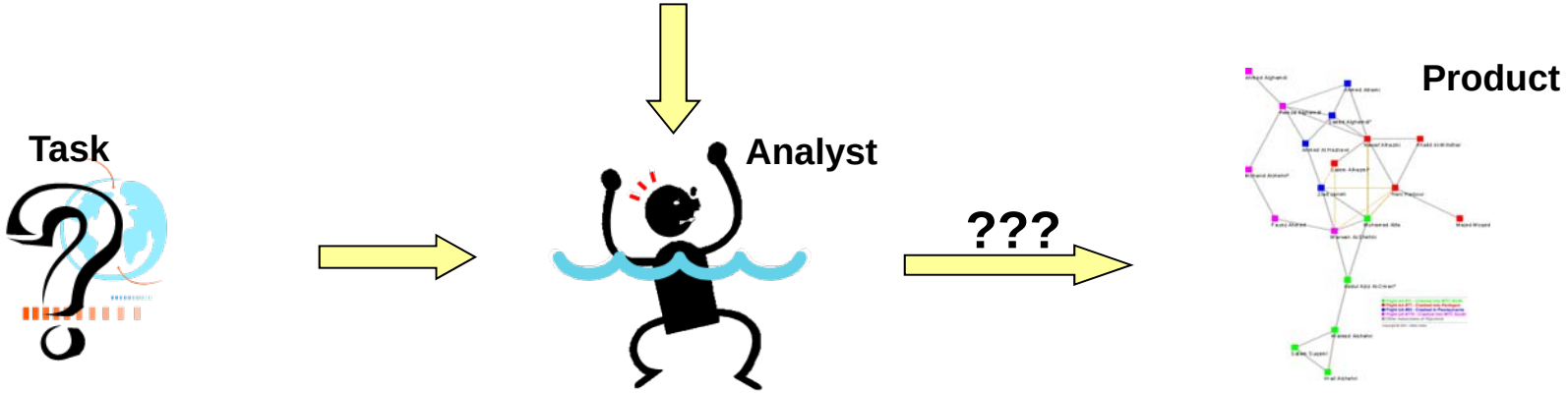
Complexity

- Complex Operations
  - Recruiting, Training, Financing, Communication, Planning, Reconnaissance, Execution
  - Covert
  - Across international borders
- Complex Data
  - Heterogeneous, un/structured, multi-source
  - Noisy, low signal
  - Very large scale, complex networks
  - Temporal, dynamic, geo-spatial
  - Many types of entities: people, places, objects, organizations, plans, events, actions, etc.
  - Many types of relations: family, affiliation, business, organizational, location, time, etc.

Data Overload



- Lots of data available
  - More data should be better – shouldn't it?
- But
  - Where to start?
  - What is relevant?



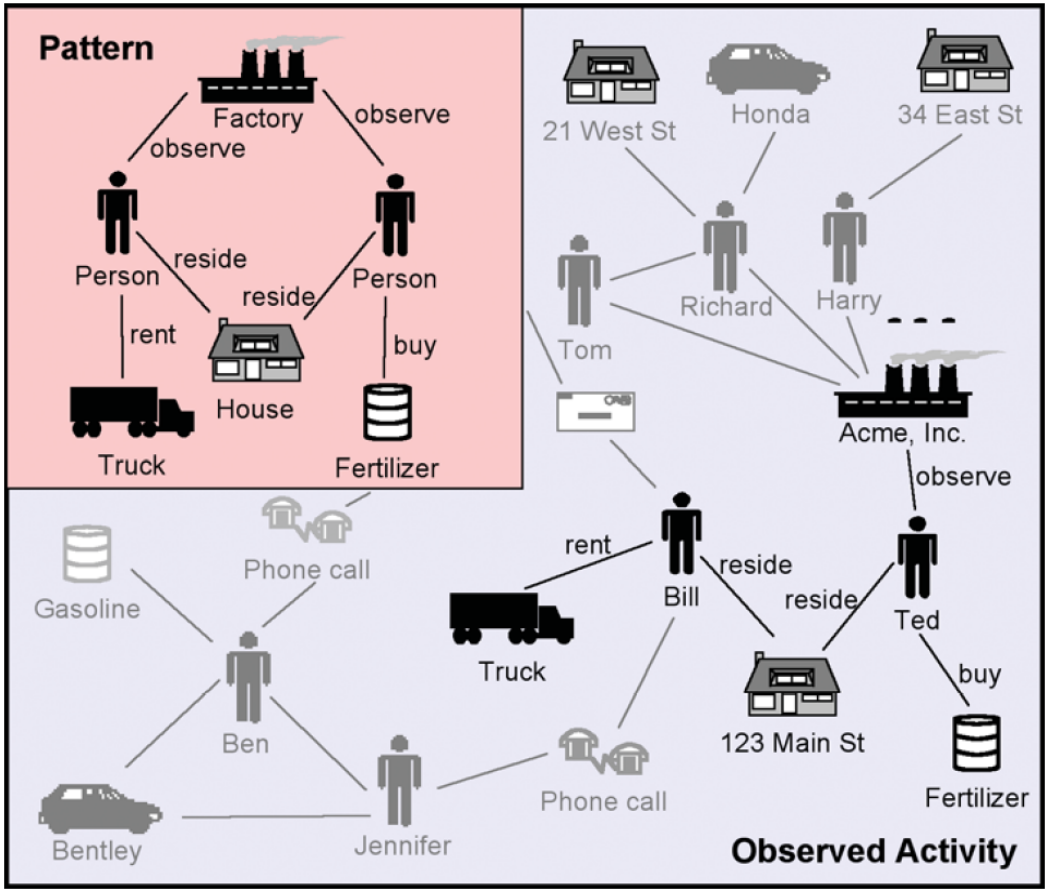
Data Analytics Tool that Can Help

- KOJAK UNICORN
  - Anomaly detection in semantic graphs
  - Detect abnormal “interesting” entities to seed tools or analysts
  - Unsupervised and unbiased
  - Integrated into IARPA’s Blackbook software (v2.8R)

- KOJAK SIMPLIFIER
  - Graph simplification to reduce data overload
  - Simplify complex semantic graphs to support visualization and discovery of interesting phenomena
  - Integrated into IARPA’s Blackbook software (v2.8R)

- P-Track (with A. Galstyan, PI)
  - Probabilistic tracking of hidden internal state (e.g., plans, intent) of potential terrorists in large dynamic agent populations
  - Current focus: modeling and tracking of Muslim radicalization

Semantic Graphs



Excerpt from: **T. Coffman, S. Greenblatt & S. Marcus, "Graph-based Technologies for Intelligence Analysis", CACM 47:3, pp. 45-47, 2004**

- Motivation
  - Link charts drawn by analysts
  - Represent/organize info about
    - People
    - Organizations
    - Places, events, etc.
    - **Relationships**
- Representation Features
  - Simple, intuitive
  - Ambiguities resolved by human K
  - Exposes multi-step connections of interest
  - Fairly easily computerized
    - E.g., store in (R)DBMS
    - Use DB queries to find instances of patterns
    - Example systems: TMODS, LAW, KOJAK, etc.

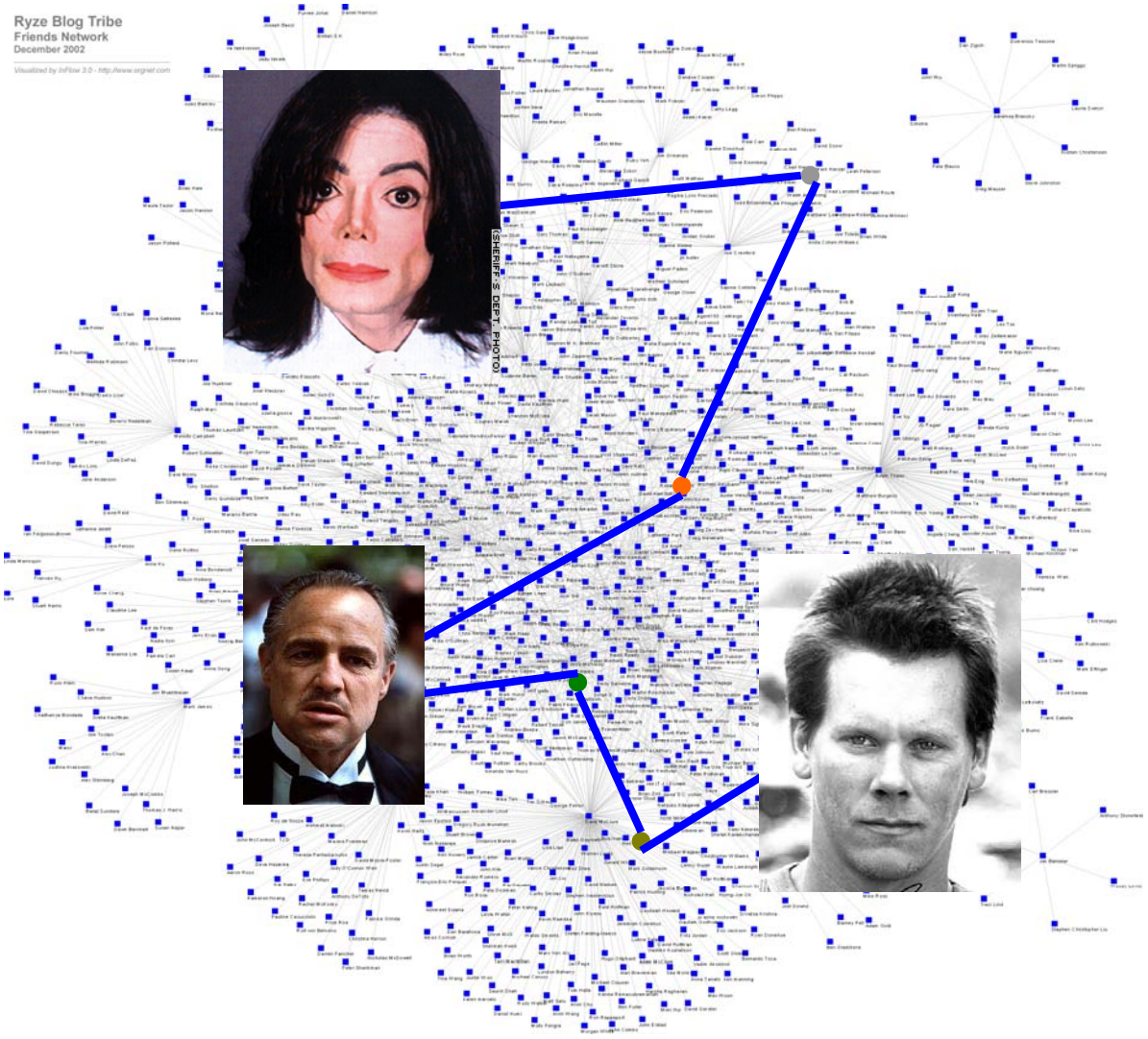
UNICORN

USC VITERBI SCHOOL OF ENGINEERING

KOJAK UNICORN

ISI Information Sciences Institute

Finding Abnormal Entities & Connections

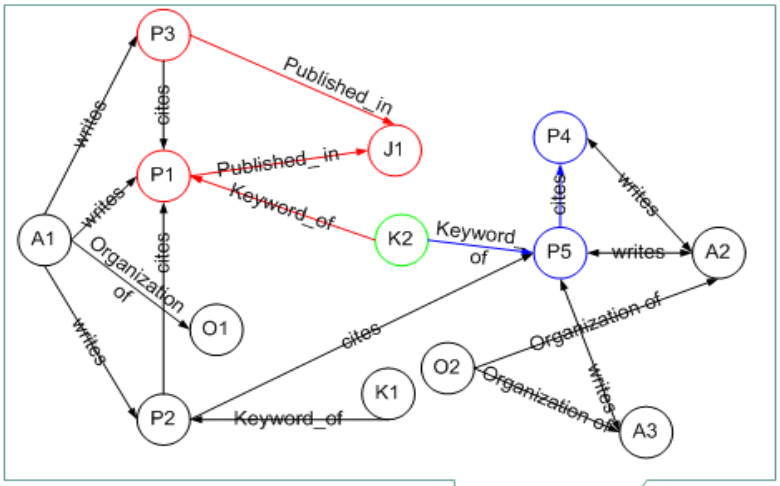


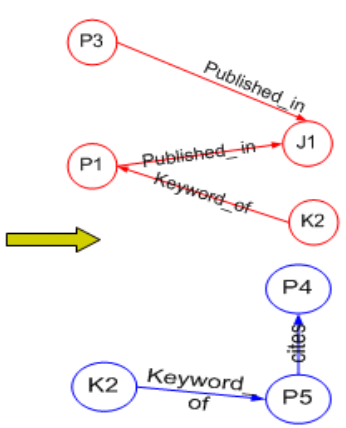
Social Network from Tribe.net

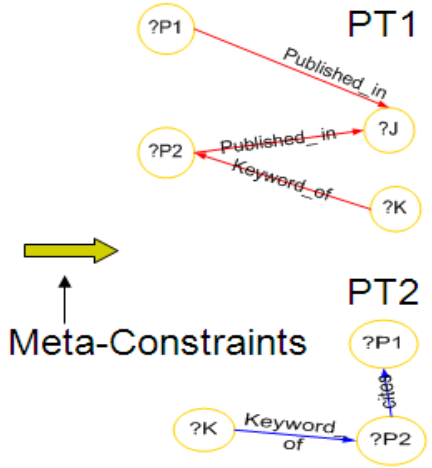
The Abnormality Hypothesis

- Why is abnormality relevant to threat detection?
  - Threat behavior is unusual, infrequent, covert, abnormal
  - Threat individuals behave in unusual ways
    - to defeat or violate the law and avoid prosecution
    - they try to fit in but get things wrong
  - CLAIM 1: such behavior will lead to unusual, “interesting” and abnormal evidence.
  - CLAIM 2: identifying abnormal entities, behaviors or connections can lead to identification of threat individuals
- Challenges
  - What does it mean for an entity or link to be abnormal?
  - How can we find abnormal, interesting behavior without knowing what we are looking for?

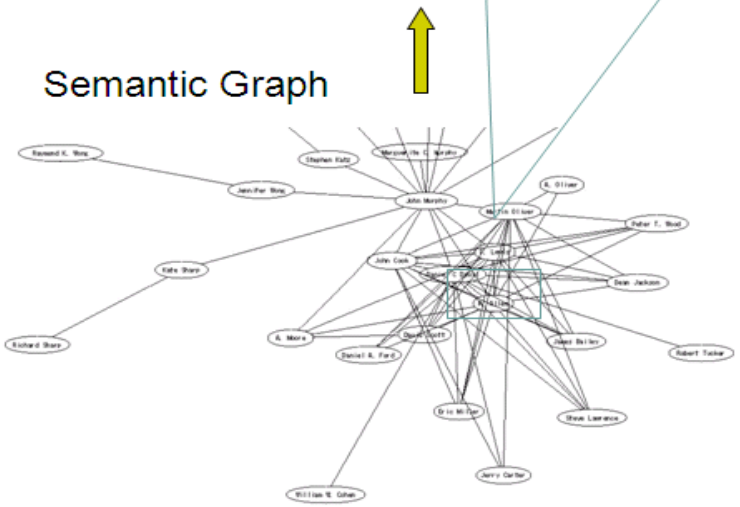
Identifying Abnormal Nodes via Semantic Profiles







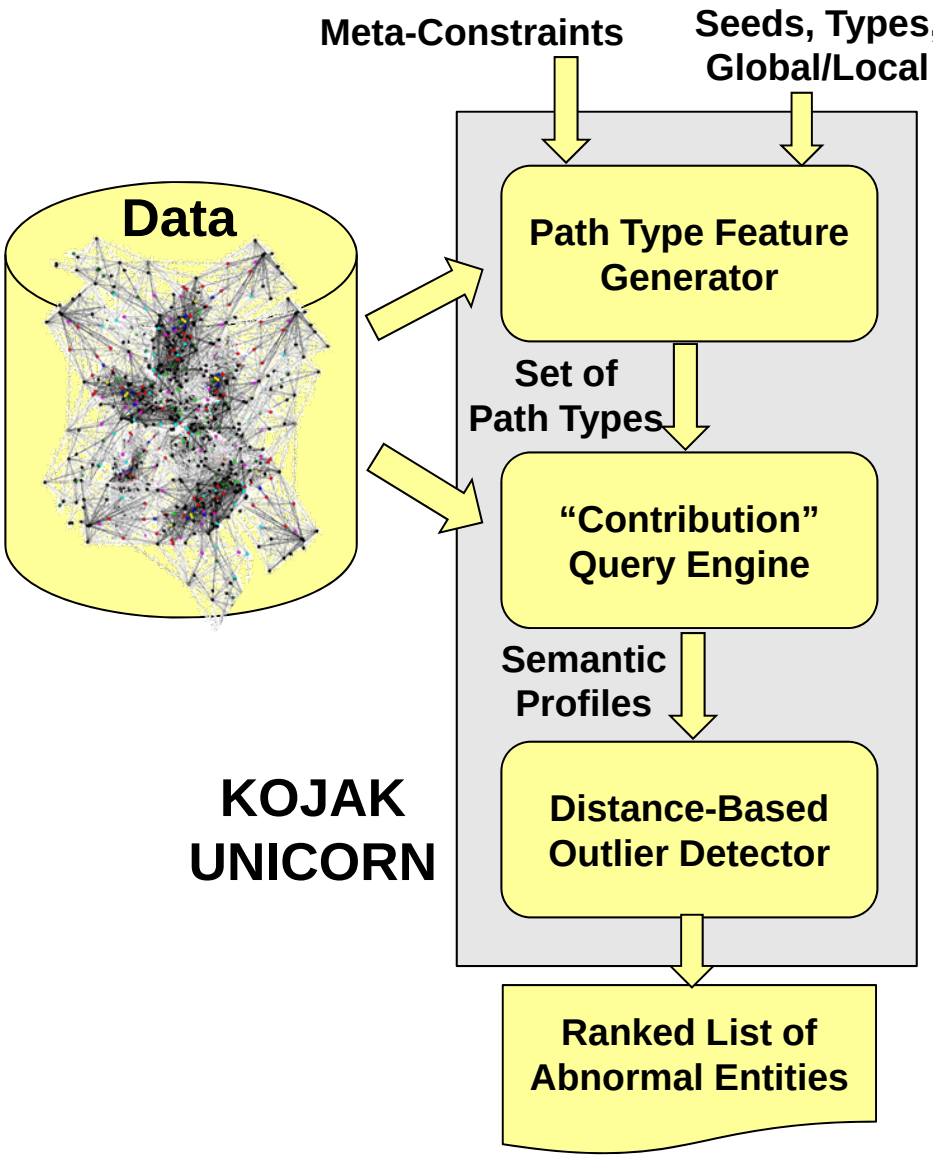
Meta-Constraints



Semantic Profile for K2

	PT1	PT2	PT3	.....
K2	50%	2%	33%	.....

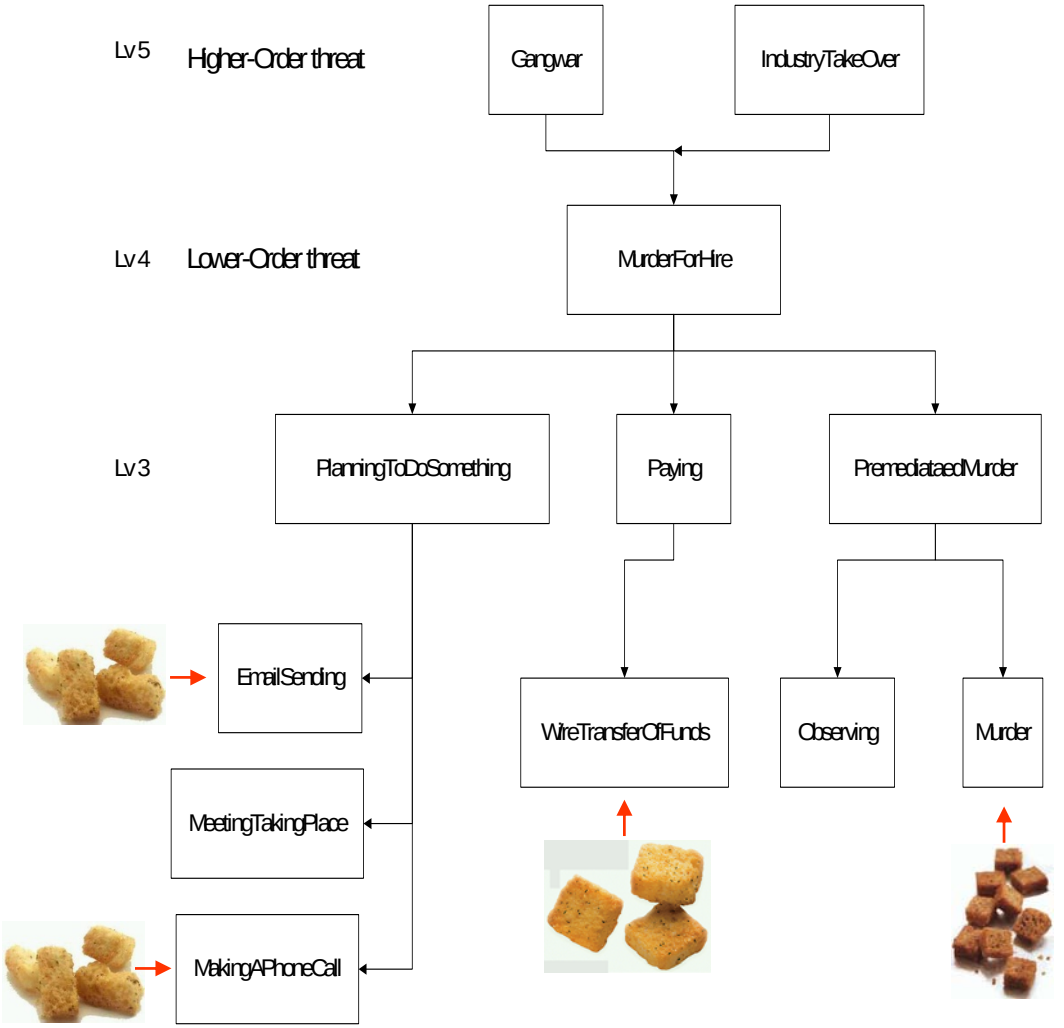
Anomaly Detection



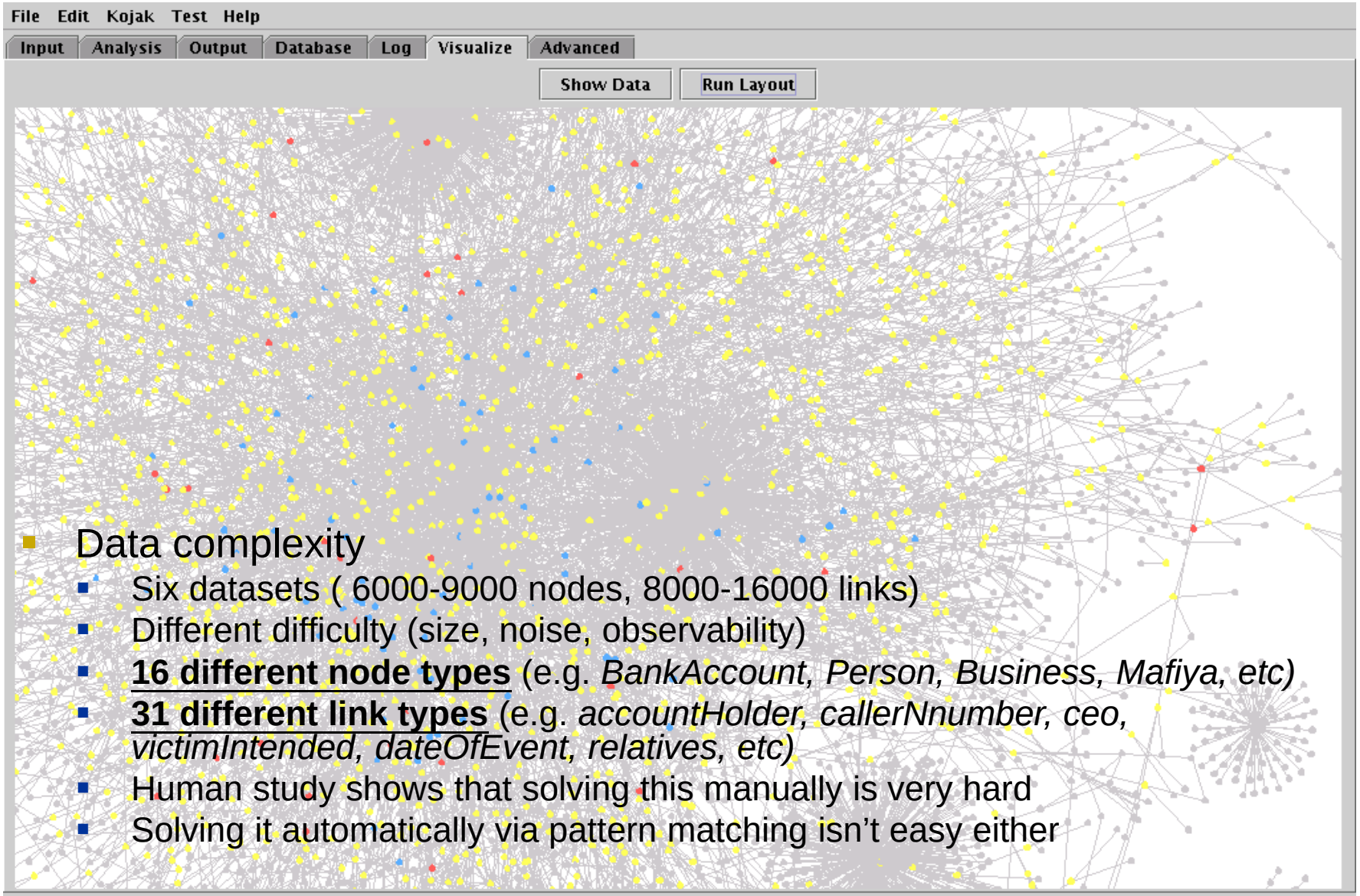
- Queries
  - **Global**: Which are the top *k* abnormal nodes?
  - **Local**: Which are the top *k* abnormal nodes connected to a given node?
- Process
  - Summarize graph neighborhood around node into **semantic profile**
    - Automatically enumerate path type features that occur in data
    - Guided via very general meta-constraints (e.g., loop free, $d<5$)
  - Use outlier detector to find nodes with abnormal semantic profiles
- Features
  - Unbiased, data driven
  - Context sensitive semantics
  - Path types have semantic interpretation
  - Results can be explained

Example Domain: Simulated Russian Organized Crime

- ROC Domain
  - 5-level task hierarchy executed by simulator
  - Low-level actions generate partial & noisy evidence
    - phone calls, money transfers, observations,…
- Task
  - Match patterns against low level evidence
  - Detect high-level events & subevent structure
- Evaluation
  - Score reported events against ground truth

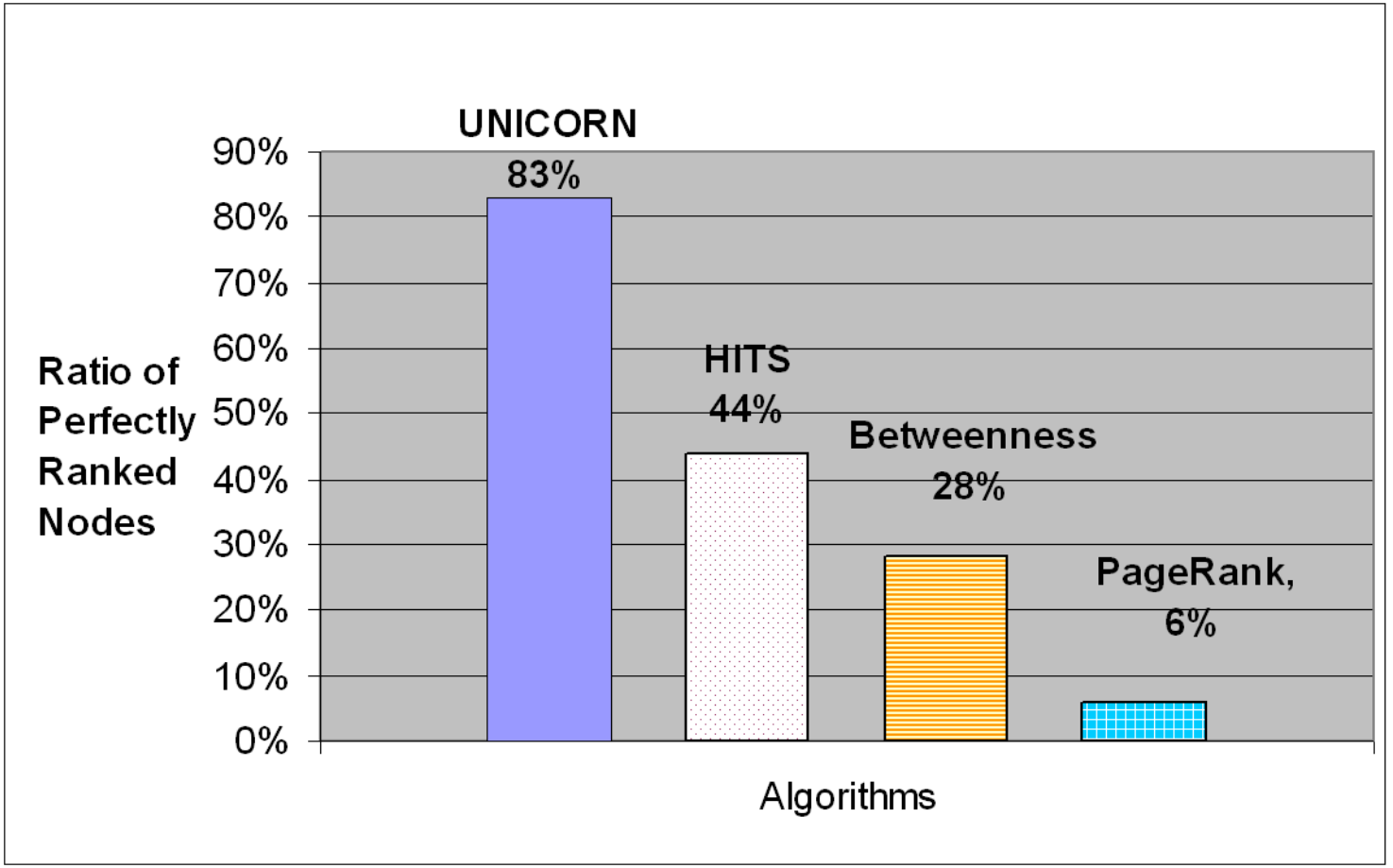


Data Complexity



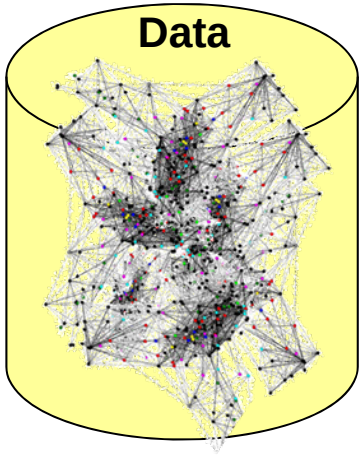
- Data complexity
  - Six datasets (6000-9000 nodes, 8000-16000 links)
  - Different difficulty (size, noise, observability)
  - **16 different node types** (e.g. *BankAccount, Person, Business, Mafiya, etc)*
  - **31 different link types** (e.g. *accountHolder, callerNnumber, ceo, victimIntended, dateOfEvent, relatives, etc)*
  - Human study shows that solving this manually is very hard
  - Solving it automatically via pattern matching isn't easy either

Comparing Different Graph Algorithms

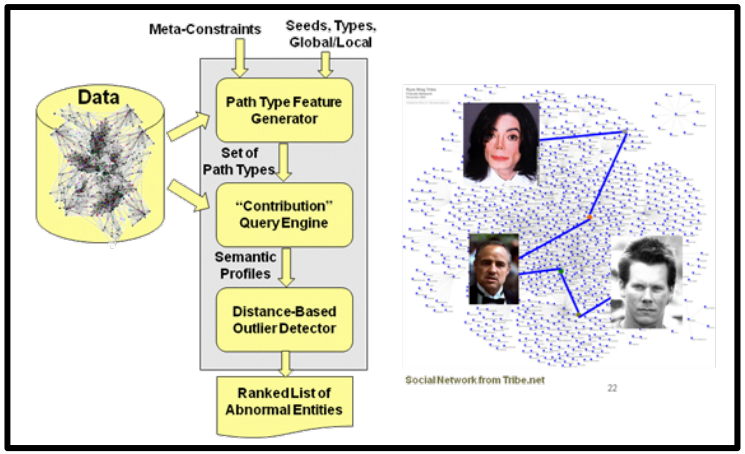


Anomaly Detection on Spinosaurus Data Reveals Threat Actors

Spinosaurus
"Border Bouncer"



KOJAK UNICORN



Top-10 Anomalous Nodes

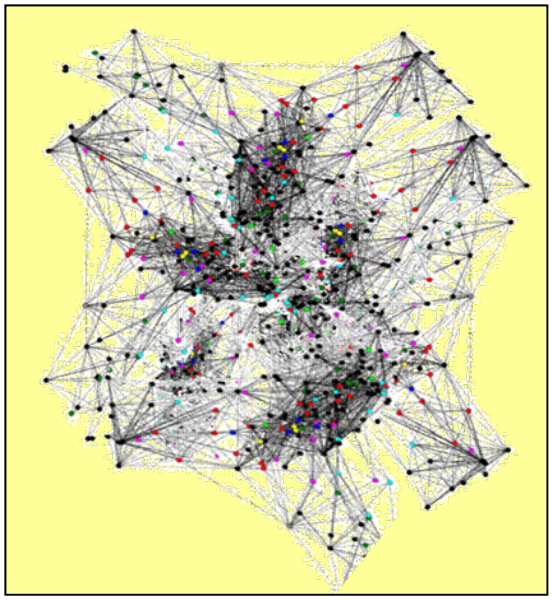
1. **Muhammed Muhammed**
2. **Mohammed Mohammed**
3. **Muhamed Amuhamed**
4. **Joseph Terlov**
5. **Bruch Terlouw**
6. **Rodman,Mitsuhiko**
7. **Cheeseman,Yonghuai**
8. **Hosaka,Rafael**
9. **Mihok,Oris**
10. **Azarbod,Tomaoo**

Elapsed Time: 2 hours

- Synthetic Data from PNNL's Threat Stream Data Generator
  - 1,000,000 records of a simulated TIC ("Terrorism Identification Center") database
  - Synthetic background with ~20 hand-inserted threat individuals
  - Developed in other DHS program to test/challenge visual analytics tools ("**tools had a hard time**", personal communication)
  - Each record describes border crossing w/ name, citizenship, departure/arrival location, group membership, watch list matches, etc.
- Applied UNICORN to find top-10 anomalous nodes
  - Converted database into semantic graph with 7 link types (arrival city, group membership, etc.)
  - **1,000,000 nodes, 7,000,000 links**
  - **Results very encouraging, top-5 are all threat people**
    - Top-3 identify different aliases of **Muhammed Muhammed**
    - 4 and 5 are the **Terlov** brothers
    - None of the Benin travelers (yet)
  - Ongoing: going to fully out-of-core processing (currently only outlier detection is out-of-core)

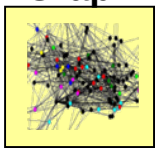
Simplifier

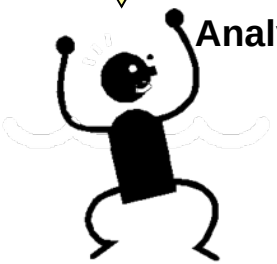
Graph & Relationship Simplification



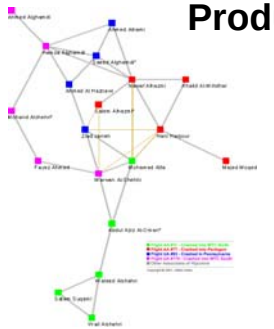
???

Simplified Graph





???



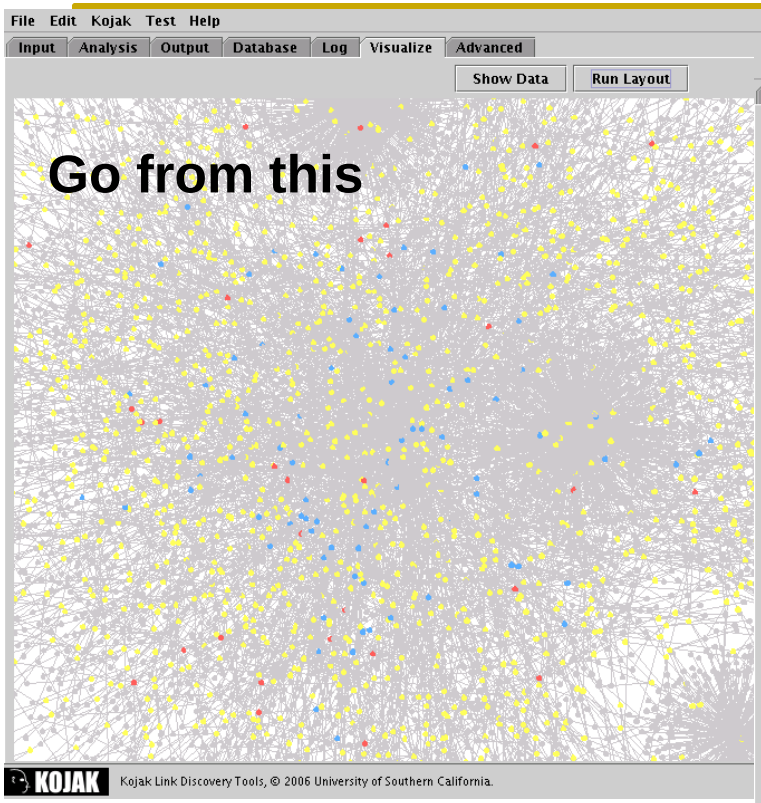
- Process
  - Normalize and abstract
  - Detect strong relations
  - Detect abnormal relations
  - Eliminate weak connections
- Features
  - Finds starting points
  - Focuses on important information
  - Context sensitive semantics
  - Reduces graph size (90% plus)
- Cost
  - Will keep some irrelevant stuff
  - Loses some information (15-20%)

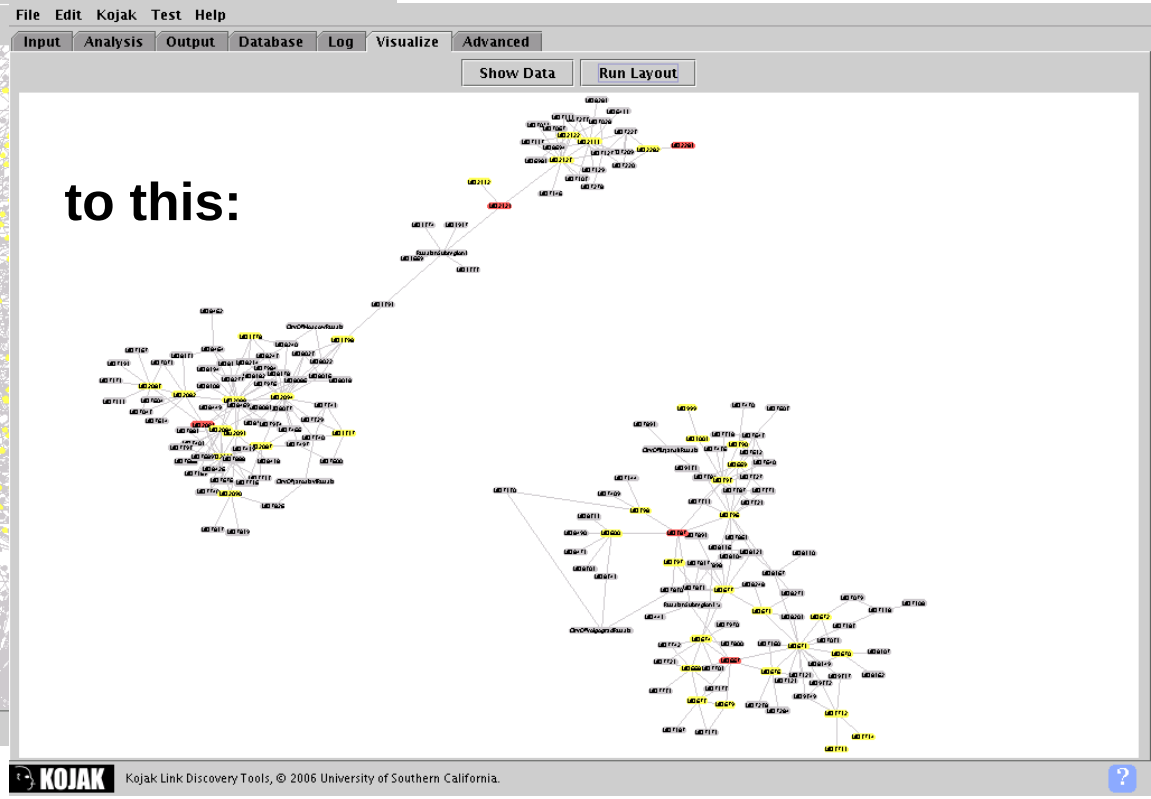
Abnormality-based Simplification

- Use abnormality as a proxy for "interestingness"

- Simplify graph and relations via
  - Focus on abnormal/interesting nodes (global)
    - Useful as a seed generator
  - Focus on abnormal/interesting nodes connected to a source/seed (local)
  - Combination of global and local simplification
  - Focus on relations most significantly contributing to the abnormality of a node

Graph Simplification on ROC Data



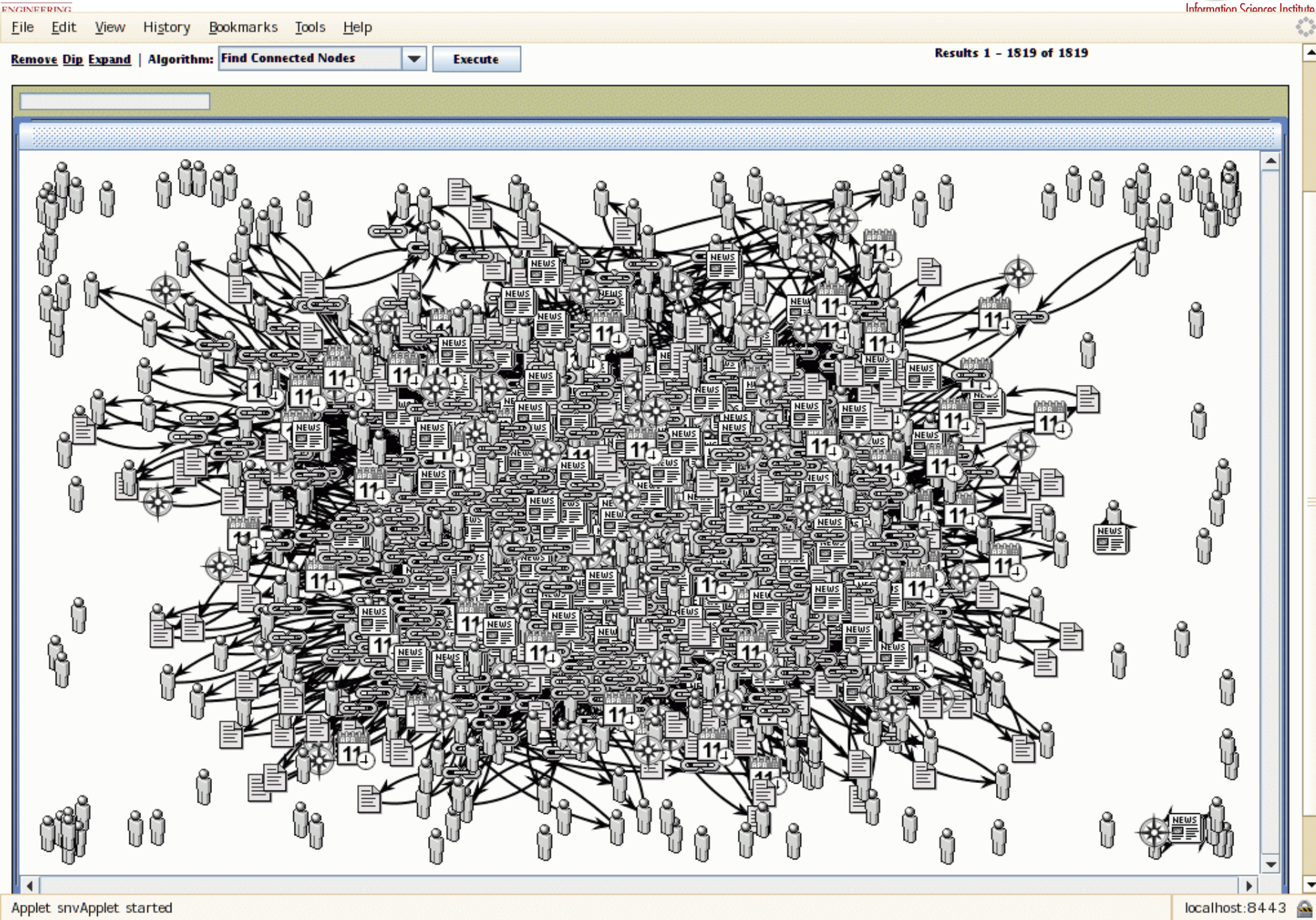


- Simplify complex semantic graphs via
  - Logic-based normalization, abstraction
  - Statistical detection of relevant nodes & links **based on abnormality**
- Preserve the essence of the data
- **Supports visualization and discovery of interesting phenomena**
- Helps intelligence analysts, link discovery tools, etc. in situations of data overload

Simplifying the 9/11 Report

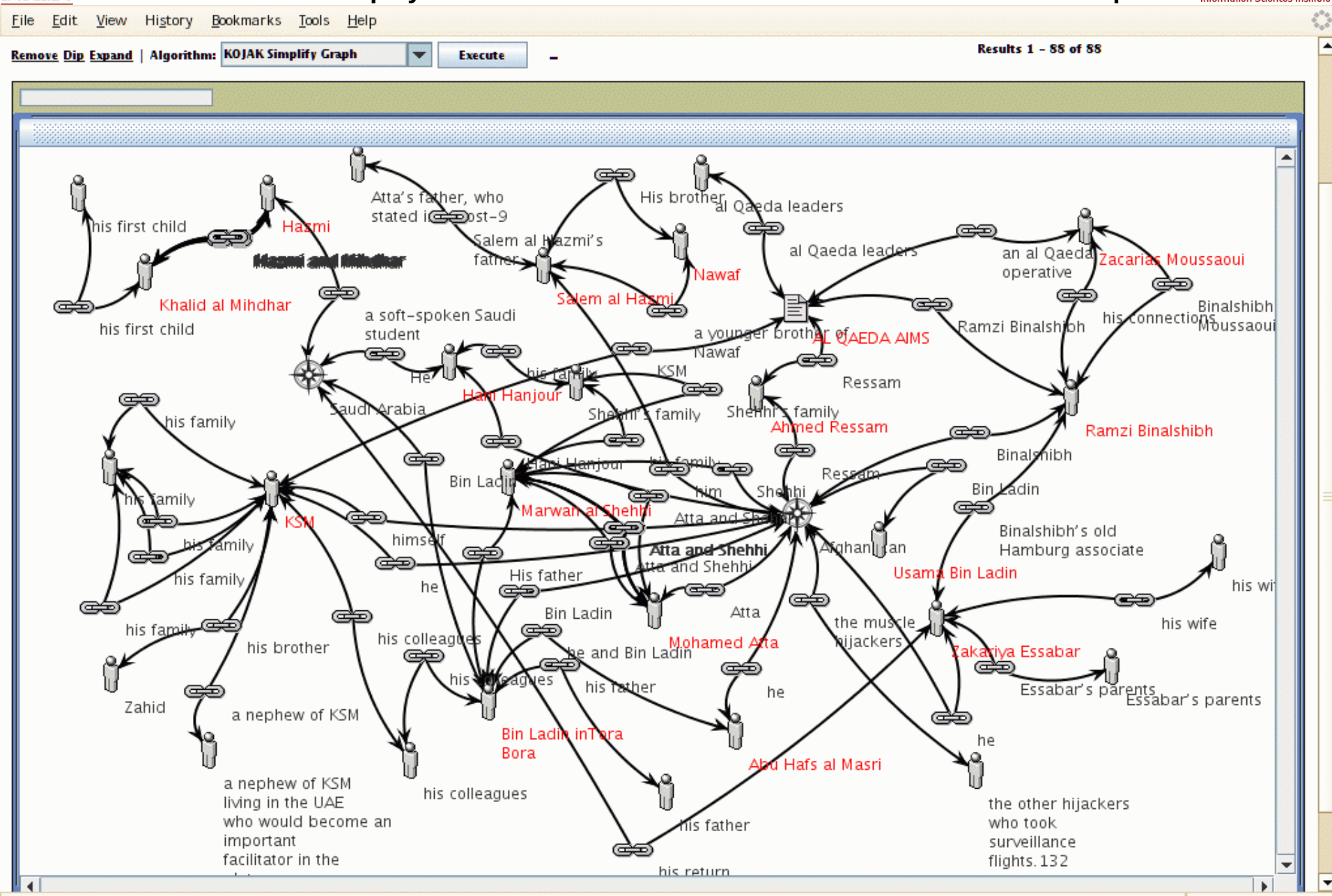
- Data: The 9/11 Commission Report, chapters 5-8, pp. 145-277
  - www.9-11commission.gov/report/911Report.pdf
  - Automatically extracted into RDF by NetOwl
    - **~100 entity types, ~300 relationship types**
    - ~5,500 entities, ~1,300 person nodes, ~2,800 links, ~40,000 triples total
  - Imported into Blackbook 2 (Analysis GUI/Middleware developed by IARPA's KDD program)
- Data slice used in experiment:
  - About 1/3rd of complete data, ~1,800 entities, ~400 person nodes, generated by querying for person nodes and then expanding/growing the graph multiple times
- Full data slice is difficult to comprehend, navigate and visualize
- After applying KOJAK Simplifier essential structure and main players in 9/11 attacks are revealed

Unprocessed data slice shown in Blackbook 2's SNV Applet (1800 nodes)



Simplified graph generated by KOJAK Simplifier shown in Blackbook 2's SNV applet

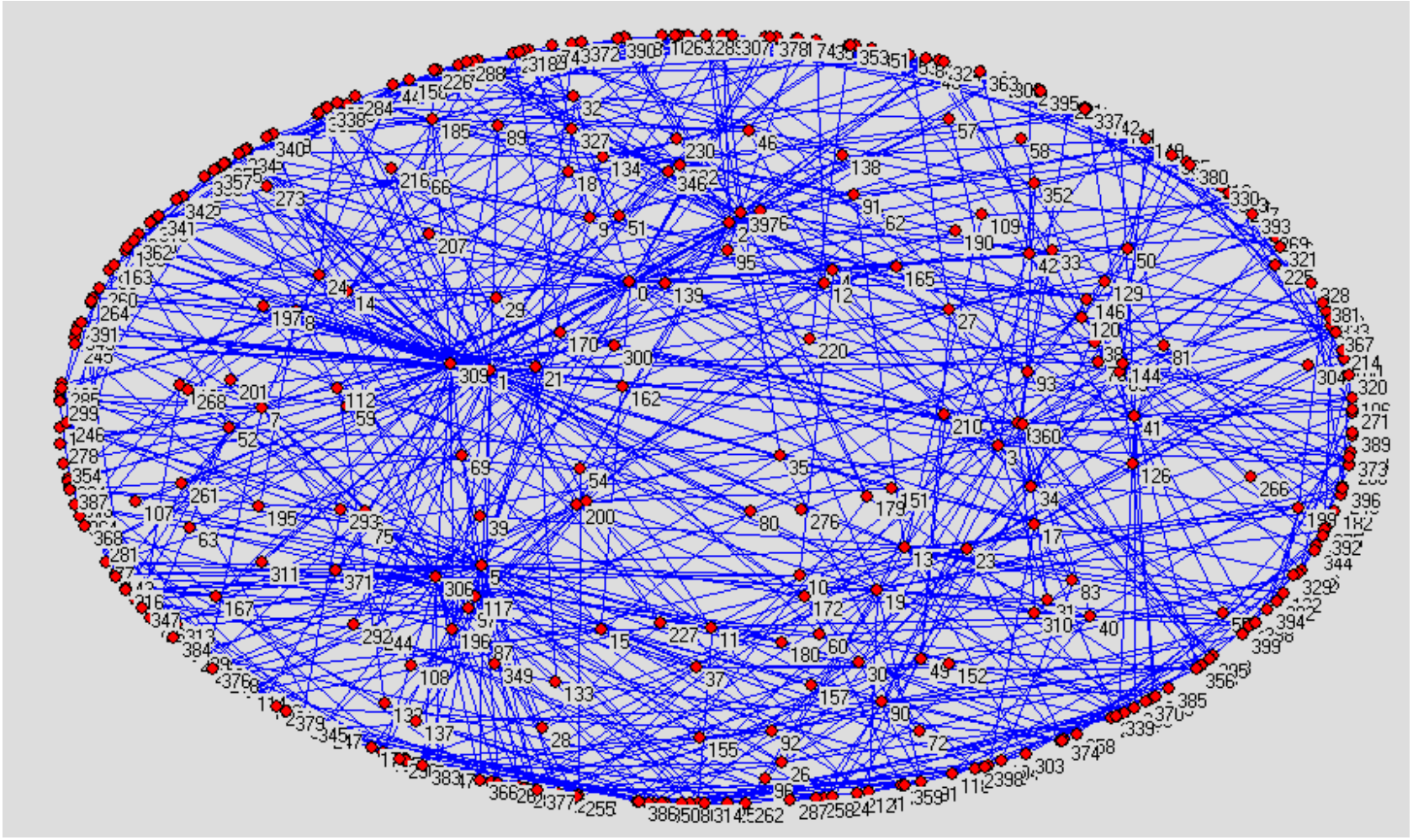
Nodes in red are core players in 9/11 attacks shown with their links and relationships



VAST-08 Phone Mini Challenge

Original data shown in Pajek, 400 nodes, ~900 multi-links, ~10,000 individual links

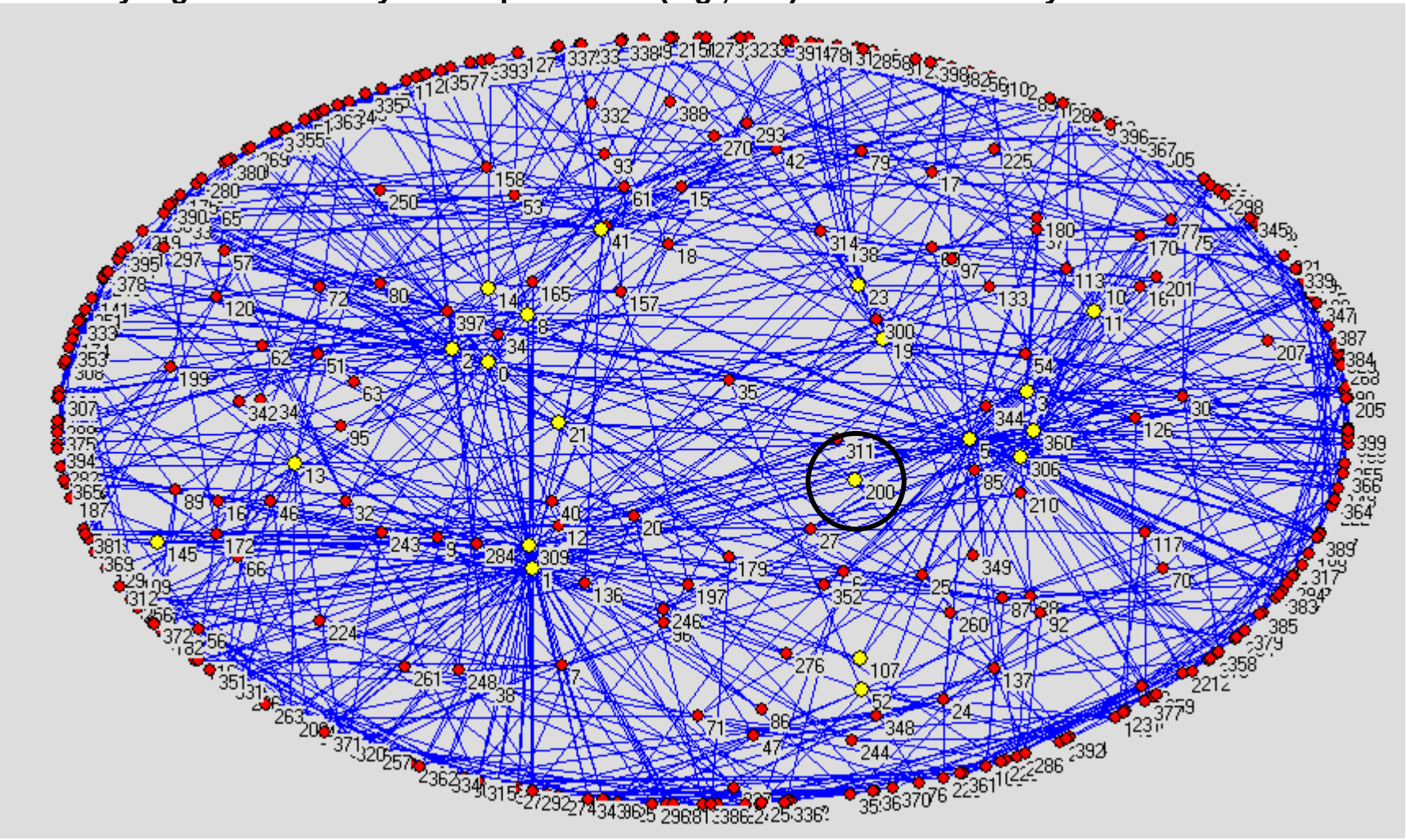
Even though dataset is small, full visualization is still difficult to make sense of



Top-20 Abnormal Nodes (UNICORN)

Abnormal nodes found by UNICORN (manually marked up in yellow in Pajek layout)

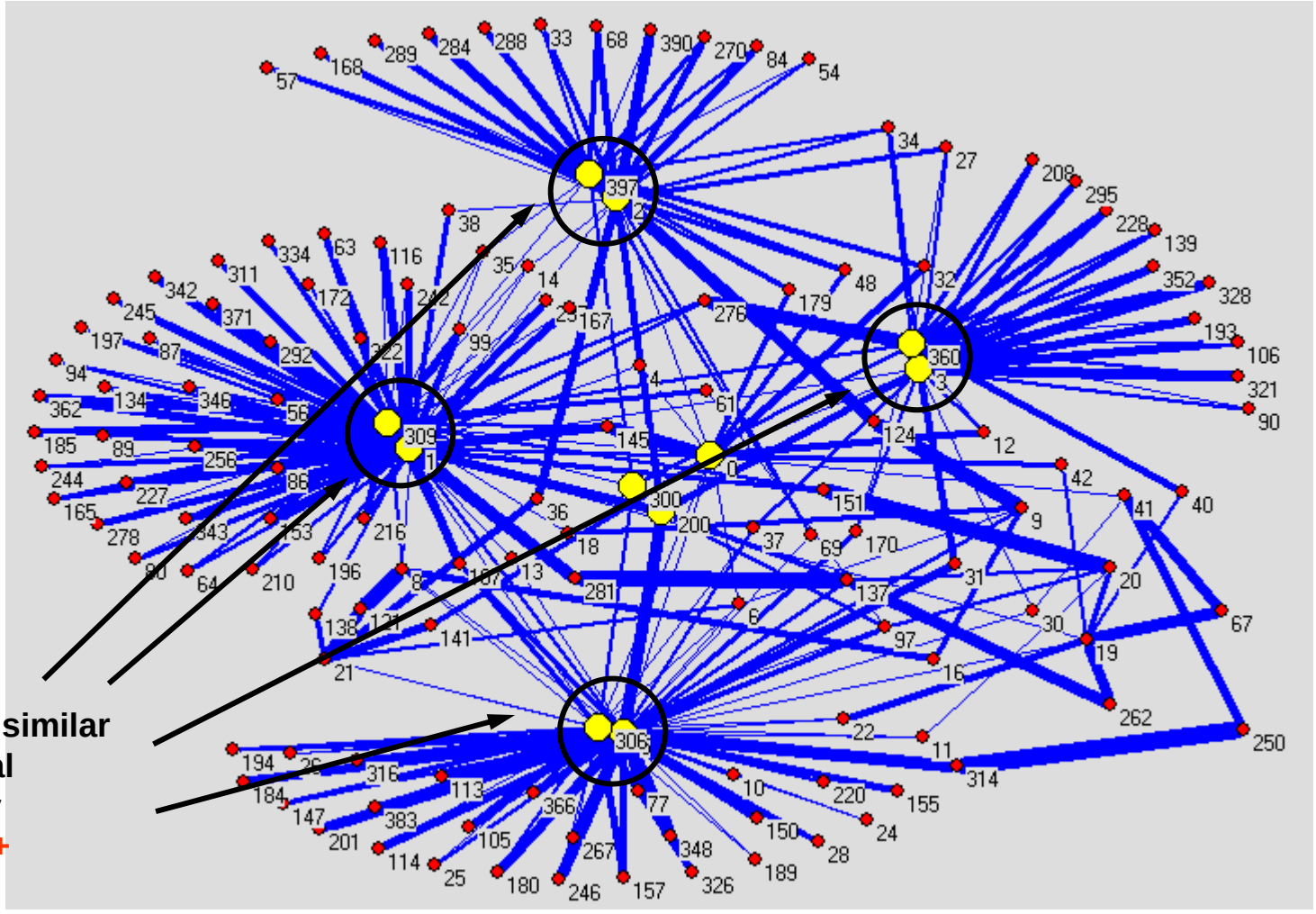
Mostly high connectivity nodes plus some (e.g., 200) that are indirectly linked to them



Simplified Graph (via KOJAK Simplifier)

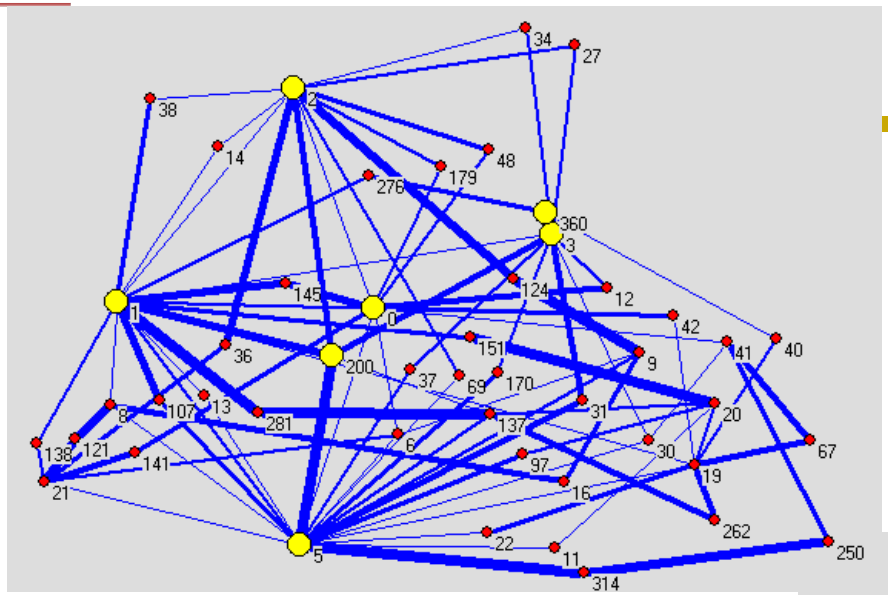
Simplified graph (all data up to Day 10) laid out in Pajek (manual color markup)

Aliases indicated by double wheel patterns seem to exist, maybe people switched cell phones?



Individuals with very similar connectivity, potential aliases uncovered by graph simplification + energy-based layout

Day 7

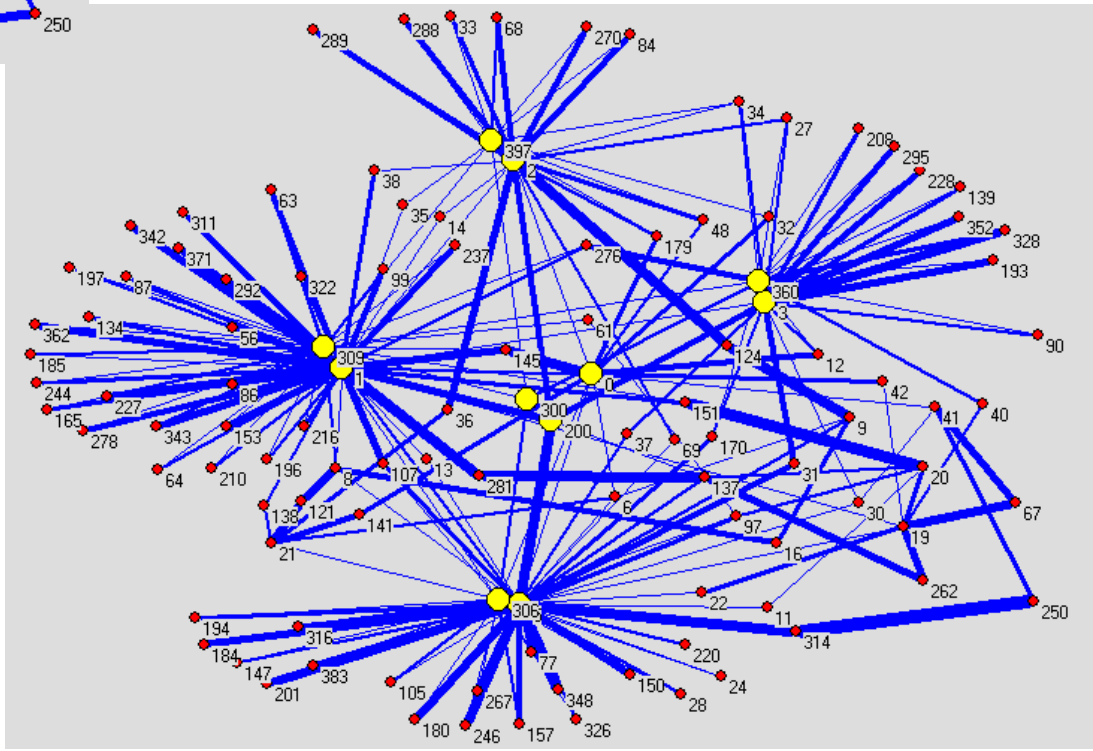


Detecting Change

Process:

- **use static Day 10 simplification and layout as a reference point**
- **then temporally animate network in reverse by eliminating links from later data**

Day 8



Temporal animation shows significant change in network structure between Day 7 and Day 8:
4 wheel patterns appear, 3 wheel hubs appear for the first time

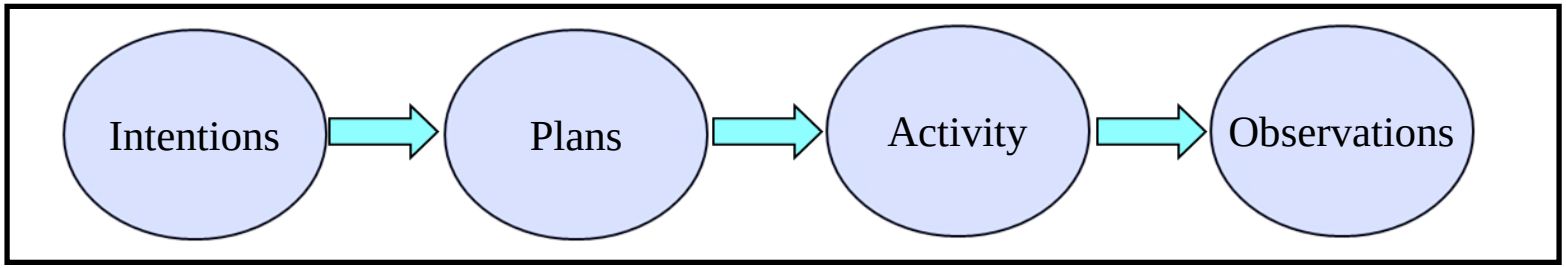
USC VITERBI SCHOOL OF ENGINEERING

ISI Information Sciences Institute

P-Track

with A. Galstyan, PI

PAIR: Plan, Activity & Intent Recognition



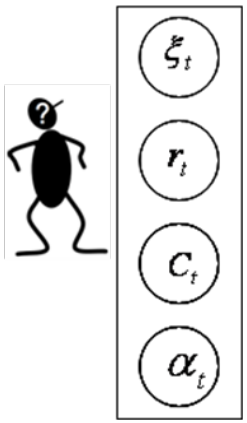
- PAIR inference problem: given a stream of observations, and some behavioral agent model that generates them:
  - What are the agents' most likely intentions?
  - What kind of a plan are they pursuing?
  - How can we estimate stages of plan execution?

- PAIR for Intelligence Analysis
  - Tens of thousands to millions of agents generating observations
  - Unknown or partially known identities of malicious agents
  - High clutter generated by a large number of agents
  - Plans involving collaborative schemes

Collective Agent Dynamics

- We model collective agent dynamics through Event-Coupled Hidden Markov Models (EC-HMMs)
  - Each individual agent is represented as a separate (hidden) Markov chain
  - Different chains are coupled through interaction events (e.g., meetings)

Individual State



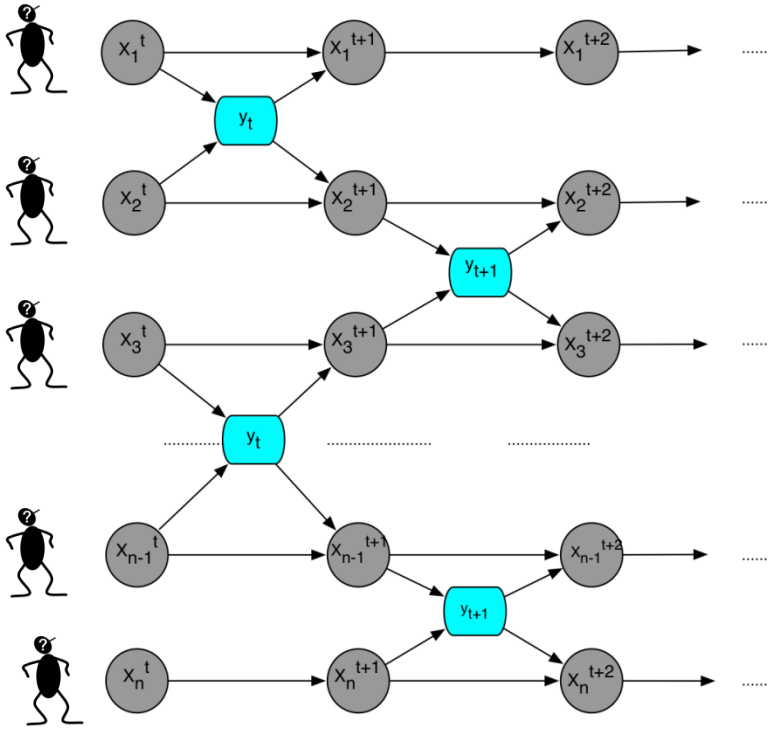
$\xi_t = \{0,1,...K\}$ Group

$r_t = \{0,1\}$ Recruiter?

$C_t = \{1,2,..M\}$ Capability

$\alpha_t = \{0,1,..M\}$ Intent

Collective State



Application: Tracking Radicalization

- **“Radicalization in the West: The Homegrown Threat”**
  - **Mitchell D. Silber and Arvin Bhatt**
    - **Senior Intelligence Analysts, NYPD Intelligence Division**

- “..While the threat from overseas remains, many of the terrorist attacks or thwarted plots against cities in Europe, Canada, Australia and the United States have been conceptualized and planned by local residents/citizens who sought to attack their country of residence.

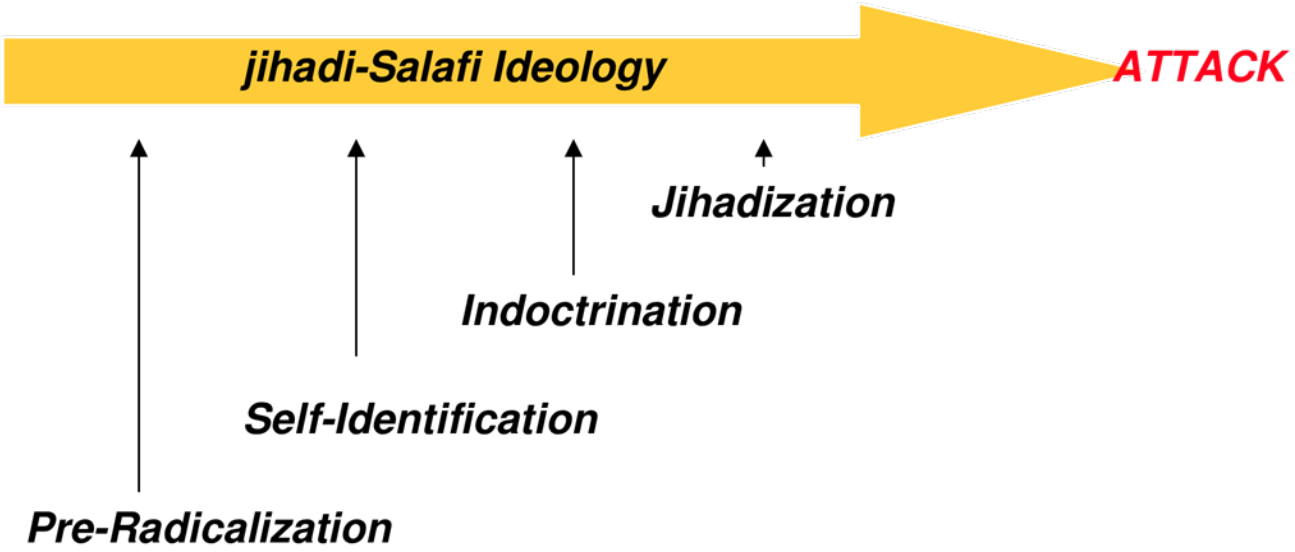
The majority of these individuals began as “unremarkable” – they had “unremarkable” jobs, had lived “unremarkable” lives and had little, if any criminal history..”

Why Focus on Radicalization?

- Our goal is to develop models for tracking terrorist activities
- Trying to detect the process at the later stages of attack planning is extremely difficult
  - Evidence is scarce
  - Little time to detect and react
- Instead, look at the "radicalization history" of potential terrorists leading up to an attack
  - It's usually a longer process (2-5 years)
- Our models are a good fit because:
  - No specific profile or a red flag that would indicate malicious intent
  - Instead, inference through accumulated evidence:
    - Need to track individual state and collective interactions
    - Need to track a dynamically evolving network

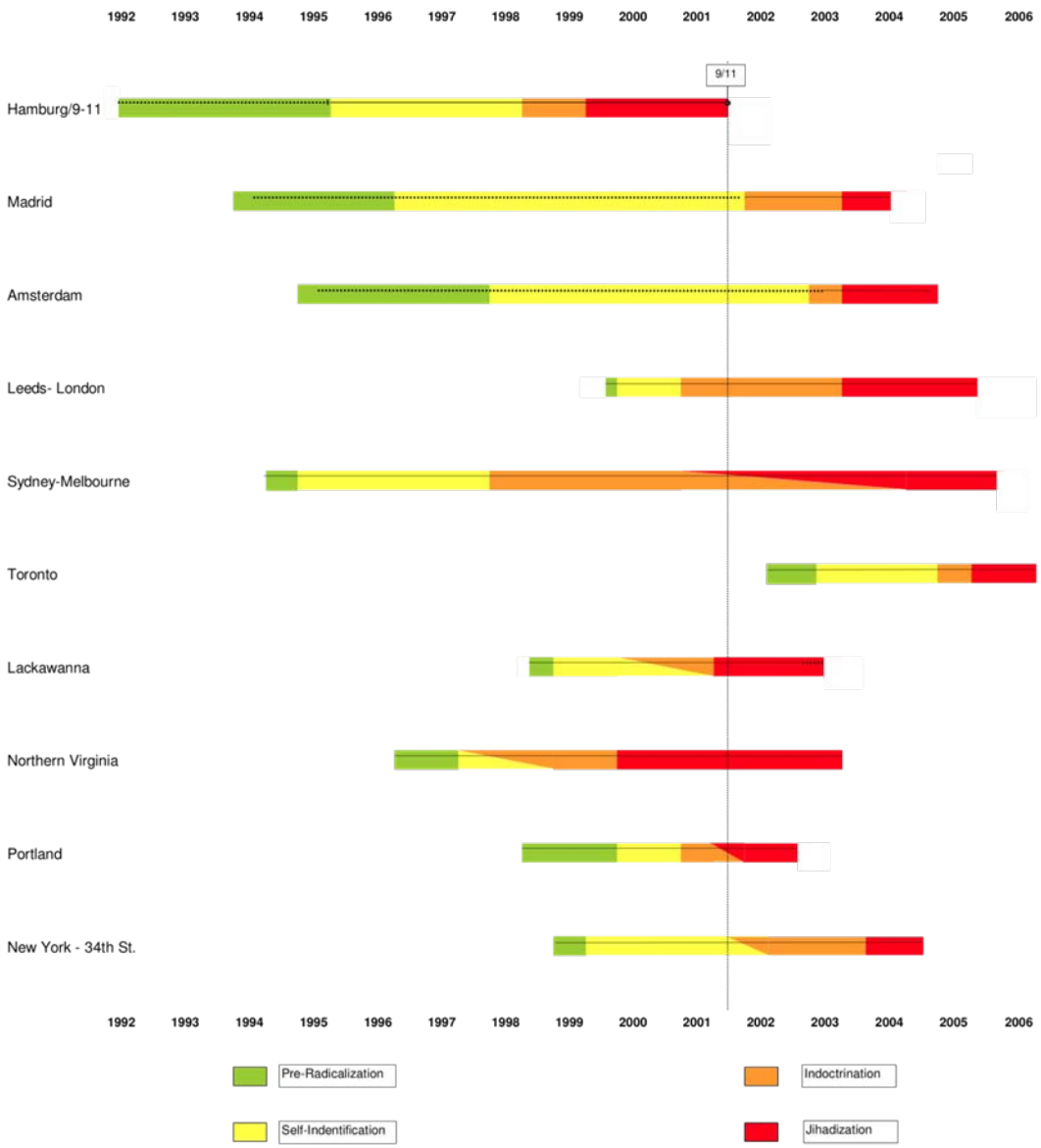
Main Findings of the NYPD Report

- An assessment of various reported radicalization models suggested that the process is composed of four distinct phases:



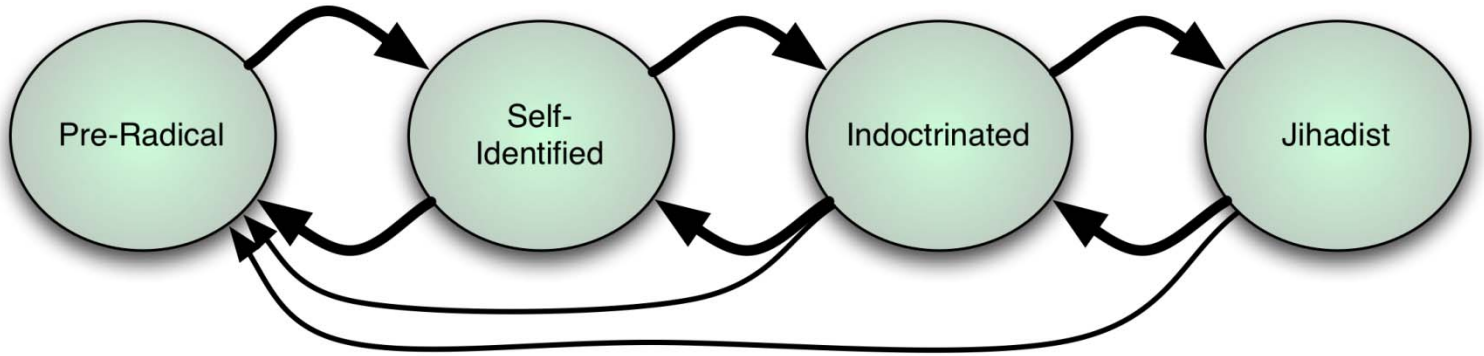
- Despite differences in both circumstances and environment in different plots, there was remarkable consistency in **behaviors** and **trajectory** across all the stages.
  - This consistency provides a tool for predictability.

Radicalization Timeline of Historic Cases



Our (Preliminary) Model

- Hidden Trajectories captured through a Markovian Model
  - Each radicalization stage is characterized as a separate hidden state in the model
  - Arrows indicate possible transitions between different stages
  - Each transition has an associated probability
    - Some transitions are more likely than others



- What type of data is available to us?
  - Reports indicating certain attributes for individuals
  - Information about a person's associations

Initial Demo

P-Track Demo

Person: B. Rubble | Load Data | Estimate State

	Observations					Estimated State			
	R	Mo	S	P	Mi	Pre-radical	Self-identified	Indoctrinated	Jihadi
Obs. 1	HI	HI	lo	lo	lo	0.460	0.358	0.130	0.052
Obs. 2	HI	HI	HI	lo	?	0.060	0.670	0.200	0.070
Obs. 3	HI	lo	HI	HI	lo	0.037	0.349	0.466	0.148
Obs. 4	HI	lo	HI	HI	lo	0.033	0.166	0.612	0.189
Obs. 5	HI	lo	?	HI	HI	0.032	0.081	0.587	0.300
Obs. 6	?	?	?	?	HI	0.077	0.090	0.272	0.561

PowerLoom 3.2.50

Conclusions

Summary: Challenges & Directions

- **Analysis is a complex, non-linear, exploratory process**
  - Hard to keep track of it all, many competing leads and hypotheses, lots of "if this is true that would follow, but…"
  - Big need for very flexible tools, hard to preconceive all situations, often need new, complex queries, new analysis tool, etc.
  - Big need for powerful hypothesis and knowledge management
- Semantic models and tools can help
  - Capture more and more aspects of a complex world
  - Sophisticated link inference, constraint checking
  - More informed similarity analysis
  - Explanation
- Future - improve this ratio:
  - 70% data understanding, translation, 5% tool application, 25% understanding results

Additional Information

- J. Adibi, H. Chalupsky, E. Melz and A. Valente (2004). The KOJAK Group Finder: Connecting the Dots via Integrated Knowledge-Based and Statistical Reasoning. In *Proceedings of the Sixteenth Innovative Applications of Artificial Intelligence Conference (IAAI-04)*
- S. Lin and H. Chalupsky (2003). Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset. *SIGKDD Explorations*, 5(2): pages 173-178
- S. Lin and H. Chalupsky (2008). Discovering and Explaining Abnormal Nodes in Semantic Graphs. *IEEE Transactions on Knowledge and Data Engineering, 20(8):* pages 1039-1052.
- H. Chalupsky and R.M. MacGregor and T.A. Russ (2006). PowerLoom Manual. USC Information Sciences Institute. Available online at http://www.isi.edu/isd/LOOM/PowerLoom

- http://www.isi.edu/~hans
- http://www.isi.edu/isd/LOOM/kojak
- http://www.isi.edu/isd/LOOM