ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

Building a Holistic Fraud Analytics Platform

Aric LaBarr, Ph.D.
Mike Thurber

Headquarters
300 W. Main Street, Suite 301
Charlottesville, VA 22903
434.973.7673 | fax 434.973.7673

Office Locations
Arlington, VA
Linthicum, MD
Raleigh, NC

About Elder Research

Elder Research delivers business value through customizable advanced analytics solutions that solve your most challenging problems.

20+ years experience

150+ customers

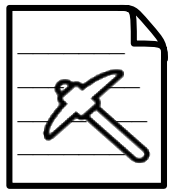
Data science experts

Trusted partner

Advanced Analytics is Our Strength

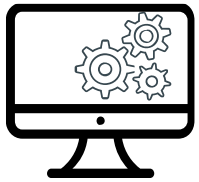
Data Science and Predictive Analytics

Discovering patterns in past data that can be used to predict the outcome of future events including statistical modeling, classification & analysis, clustering, optimization & simulation, and customer segmentation



Text Mining

Understanding information stored in text documents and databases including document classification, natural language processing, information extraction and search



Data Infrastructure

Cleaning, preparing, and integrating disparate data sources and building ETL and data pipelines optimized for advanced analytics

Data Visualization

Making advanced algorithms easily accessible through 2-D & 3-D, statistical and spatial visualization

Previous Webinars on Fraud

- Best Practices for Deploying a Fraud Analytics Solution (Jan 25)
- Detecting Fraud with Graph Approaches (March 8)

https://www.elderresearch.com/company/resource-center/webinars

Today's topics

- Steps toward a holistic mature fraud analytics platform
  - Unsupervised
  - Supervised
  - Semi-supervised
  - Unsupervised on "goods"
- Challenges
- Bias mitigation
- Graph features
- Stories!

Types of Fraud

- Bogus insurance claims
- Identity theft
- Tax filing
- Financial statements
- B2B (dealer) fraud
- Cash skimming
- Excessive medical service billing
- Worker's compensation fraud
- Pharmacy claim and medication fraud

Levels of Fraud Sophistication

- Independent perpetrator—
  - Such as exaggerating property damages or stealing a credit card
  - Very common
- Lightly networked
  - Such as “looking the other way”
- Embedded professional network scheme
- Emerging professional fraud scheme

The Fraud Problem

- Different industries have different approaches to fraud detection and problems.
- Regardless of the industry, two things are important for any fraud detection solution:
  1. Observing **known** fraudulent observations to determine patterns that may assist in finding other fraudulent observations.
  2. Observing behavior and identifying suspicious actions that might be fraudulent – lead to further investigation and identification of **new** fraudulent observations.

Levels of Fraud Analytics Maturity

Where does your organization fit?

Level	0	1	2	3	4
Enterprise Fraud Maturity	New	Young	Emerging SIU	SIU with basic fraud scoring	Holistic Fraud system

After a new organization is victimized by fraud, investigations are triggered:

1. Someone may call in a tip. Internally, someone may point out something suspicious.
2. As an organization matures, a fraud team, especially a Special Investigations Unit (SIU) will be put in place to proactively scan, search, and investigate to uncover fraud. Common sense rules and heuristics will emerge naturally from the experience of fraud investigations.
3. Eventually, the need for predictive fraud models will be recognized, and investigations will be triaged and prioritized by fraud scores.
4. Finally, the need for a holistic analytic infrastructure will be recognized and implemented to combat existing and emerging fraud types.

Levels of Fraud Analytics Maturity

What are your analytic assets?

Level	0	1	2	3	4
Enterprise Fraud Maturity / *Fraud System Component*	**New**	**Young**	**Emerging SIU**	**SIU with basic fraud scoring**	**Holistic Fraud system**
Simple rules					
Proprietary rules					
Unlabeled data					
Labeled fraud instances					
Labeled non-fraud (good) instances					
Tabular features					
Graph features					
Deployed graph systems					

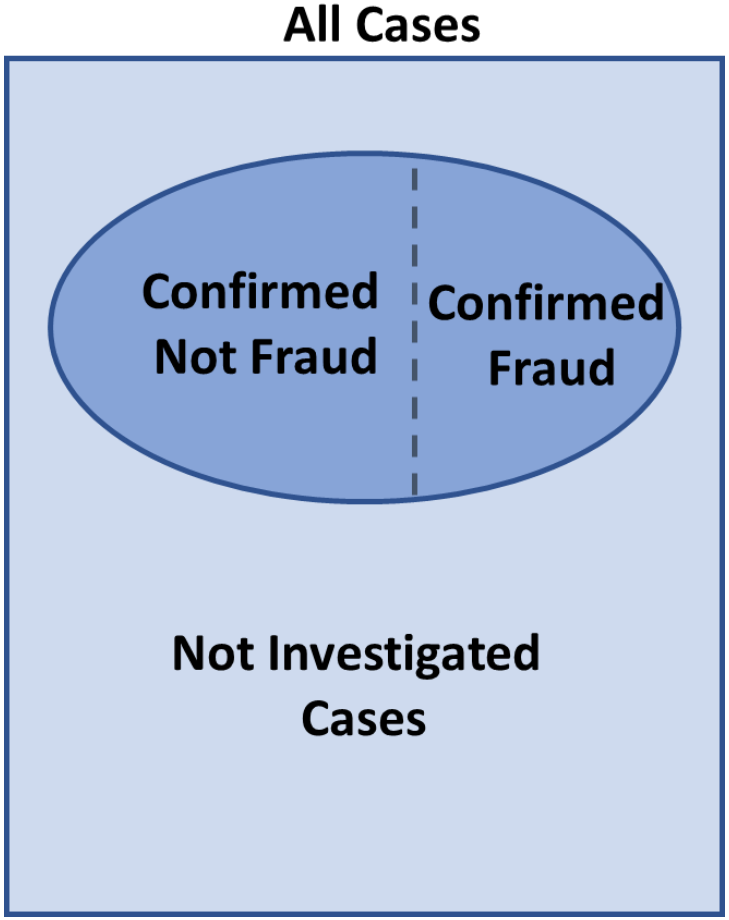
Levels of Fraud Analytics Maturity

Assets develop with maturity

Level	0	1	2	3	4
Enterprise Fraud Maturity / *Fraud System Component*	**New**	**Young**	**Emerging SIU**	**SIU with basic fraud scoring**	**Holistic Fraud system**
Simple rules	Yes	Yes	Yes	Yes	Yes
Proprietary rules	No	No	No	Yes	Yes
Unlabeled data	No	Yes/No	Yes/No	Yes/No	Yes
Labeled fraud instances	No	No	Yes	Yes	Yes
Labeled non-fraud (good) instances	No	No	No	Yes	Yes
Tabular features	Yes	Yes	Yes	Yes	Yes
Graph features	No	No	No	Yes	Yes
Deployed graph systems	No	No	No	Yes/No	Yes

Universe of Potential Fraud Cases

- Business rules or intake analyst determines which cases to send to SIU
- SIU investigates some and reports:
  1. Discovered fraud
  2. Confirmed non-fraud *(hopefully)*
- Others remain never investigated



Models for Holistic Fraud Platform

Models for Holistic Fraud Platform

0. Anomaly models for everyone
   - For Level 0 and 1 maturity
   - Anomalous ≠ Fraudulent, so SUI investigates conservatively
   - Refreshed until labeled cases begin to emerge
1. Fraud Likelihood Supervised Model
   - For levels 1-4
   - Refreshed annually
2. Not-fraud model
   - For levels 3-4
   - Refreshed quarterly or more often
3. Clusters of “not goods” model
   - For level 4
   - Refreshed often

Models for Holistic Fraud Platform

0. Anomaly models for everyone

1. Fraud Likelihood Supervised Model

2. Not-fraud model

3. Clusters of “not goods” model

Who is anomalous?

EpicLOL.com

Anomaly models for all cases generally

1. With SME knowledge, build a limited (<30) fraud detection feature set
   (Give each subject area equal number of features)
2. Build unsupervised models
   - CADE: “Which ones look least like the others?”
   - Isolation Forest: “Which ones are most isolated?”
   - Semi-supervised (very incomplete fraud labels):
     “What larger patterns do the fraud cases belong to?”
3. Apply the models to current set
4. Investigate the most suspicious cases for fraud
5. Report and label the investigated cases

Models for Holistic Fraud Platform

0. Anomaly models for everyone

1. Fraud Likelihood Supervised Model

2. Not-fraud model

3. Clusters of “not goods” model

Distinguish fraud from not-fraud

- Among investigated cases (easy)
  - Tells SIU who to investigate first
  - Does *not* indicate who to investigate
- Apply to *all* cases (hard)
  - Must be built on cases that are representative of all cases
    - Selection bias mitigation is the essential difference
  - Applies to both investigated and uninvestigated cases
  - Still, not designed to find new emerging fraud types

What Most Don't Do – But Should…

- Handle selection bias
  - Why build ONLY on investigated cases?
  - Must weight back to population properly.
  - Bootstrapping
- Build fraud likelihood model using bootstrapped sample
- Explain scores
- Investigate top cases
- Report and label the investigated cases

Creating Powerful Unsupervised Fraud Models

Models for Holistic Fraud Platform

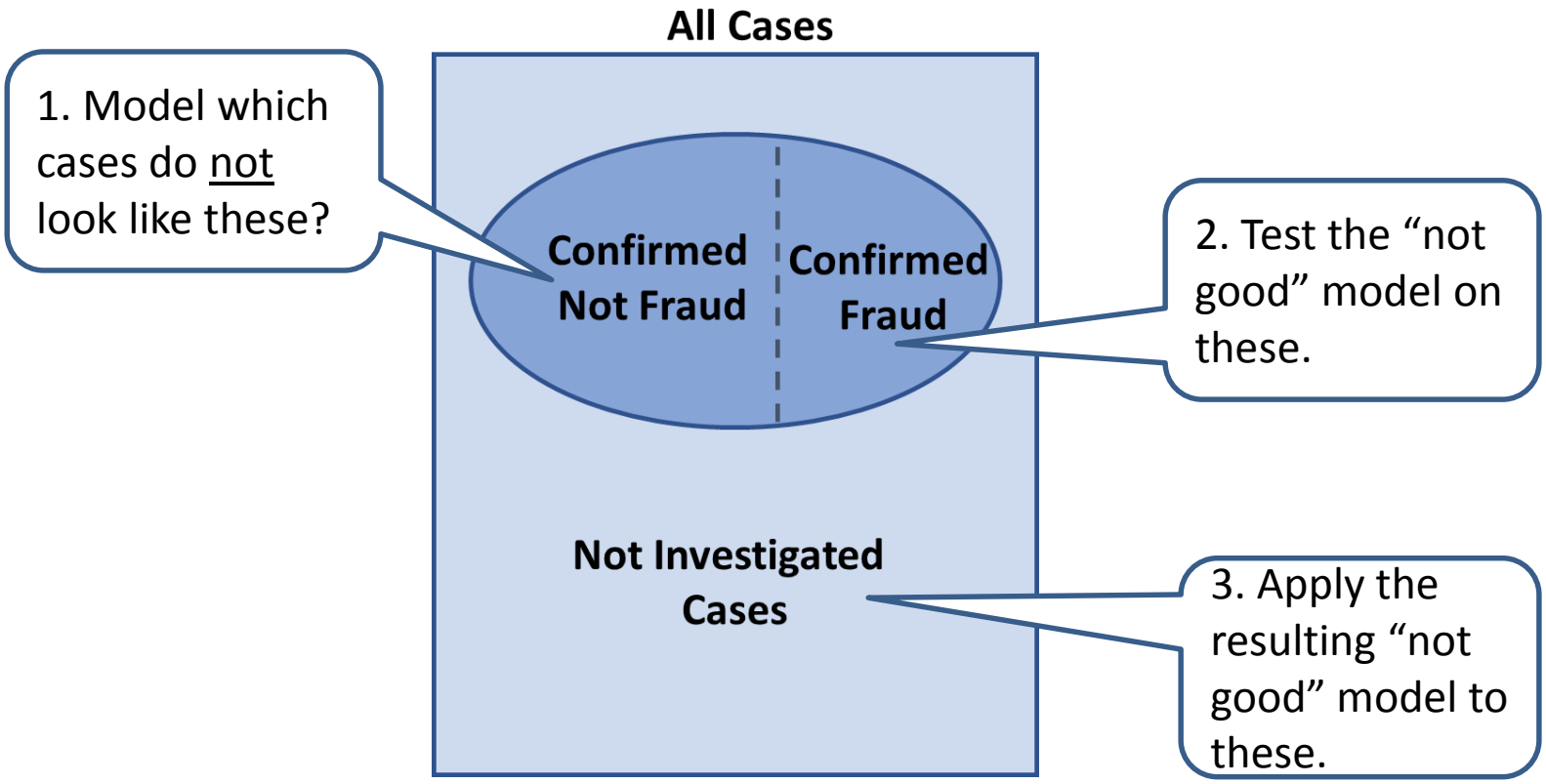
0. Anomaly models for everyone

1. Fraud Likelihood Supervised Model

2. Not-fraud model

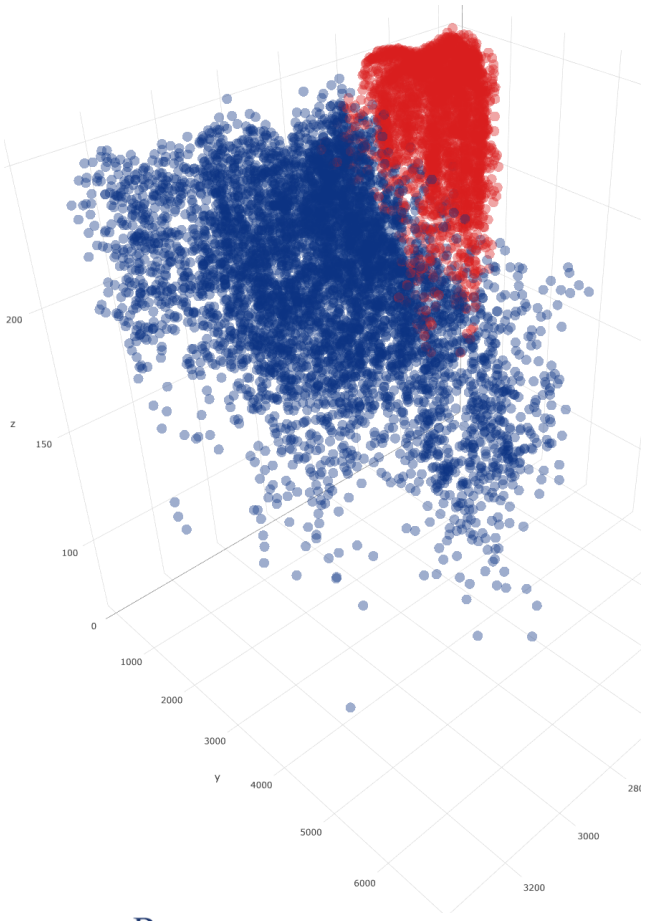
3. Clusters of “not goods” model

Who doesn't look like the known goods?

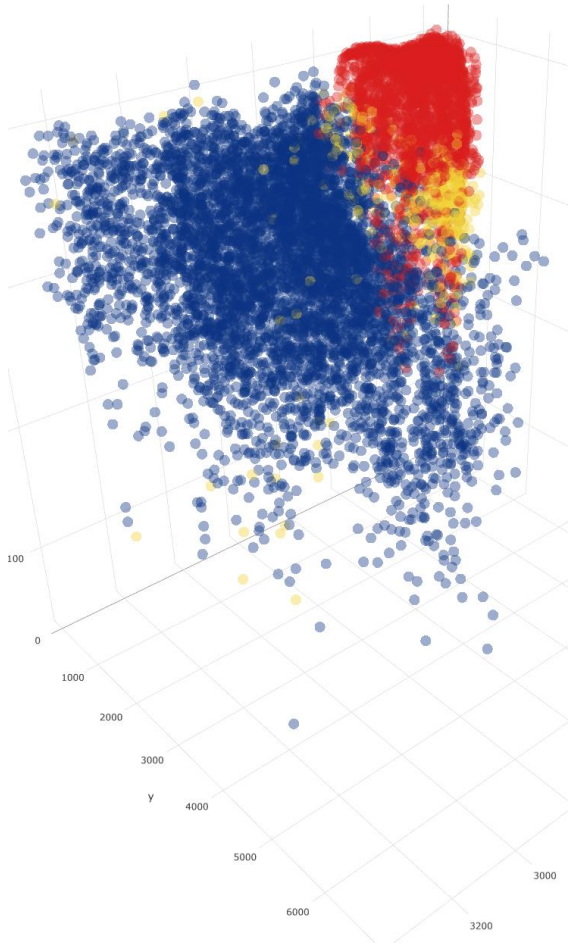


CADE Example

Supervised

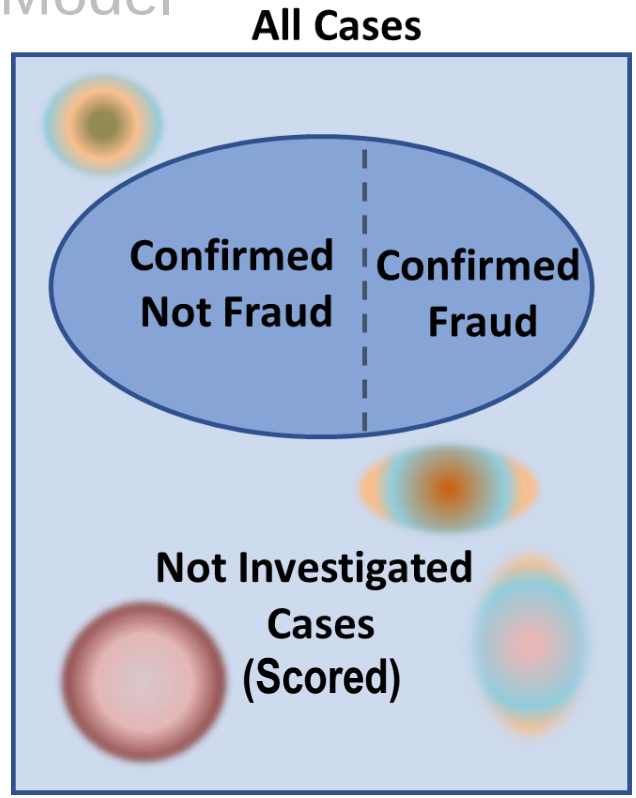


Supervised + CADE



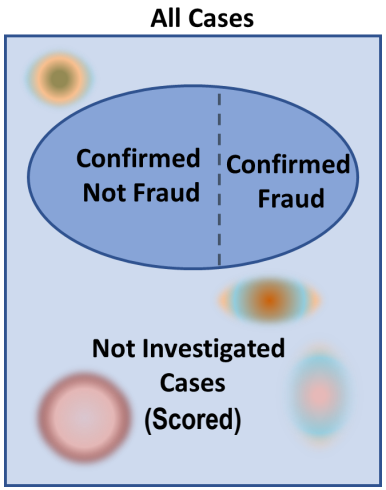
Models for Holistic Fraud Platform

0. Anomaly models for everyone
1. Fraud Likelihood Supervised Model
2. Not-fraud model
3. Clusters of “not goods” model



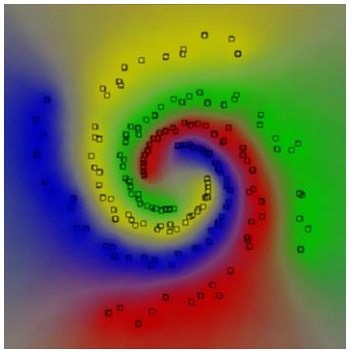
Clusters of “Not Goods” Model

1. Apply Not-fraud model to *uninvestigated* cases
2. Extract the *most likely not good* cases
3. Select the inputs used to build the fraud likelihood model
4. Apply a clustering algorithm to this set
5. Study the clusters (of the “not goods”) to decide which are most suspicious, considering
   a. Explained scores from fraud likelihood model
   b. Explained scores from “Not good” model
   c. SME input
6. Investigate the most suspicious clusters for fraud
7. Investigate the most anomalous cases
   a. Cases without any clear cluster assignment
   b. Isolation forest anomalies
8. Record the investigation case results



Many Alternatives

- Isolation Forest, like CADE, is effective at finding anomalies amid the “normals.” Caution:
  - Anomalous ≠ Fraudulent
  - Works best when 300<n<1000
- Semi-supervised techniques can be very effective, especially for truly rare events (terror attacks, regime changes)
  - Find cohesive group that labeled case belongs to
- Networked relationship features are typically very powerful—stay very open-minded!
  - Were their known bad actors in their community?
  - How many providers does the customer or patient *drive by* who are closer than their selected one?
  - Was the timing and sequence of events as expected?



Supervised vs. Unsupervised

Technique	Pros	Cons
• Unsupervised alone	• Requires no labels, finds emerging fraud	• Poor accuracy, naive
• Supervised alone	• Highly efficient, defensible	• Misses new fraud paradigms entirely
• Sequential Supervised→ Unsupervised	• Fully leverages supervised and unsupervised benefits	• Holistic fraud platform required

Case Study – Opioid Epidemic / Pharmacy Fraud

Opioid Epidemic/Pharmacy Fraud

Problem:

Figure out which provider is abusing the system?
Also cluster services (there are thousands of them)

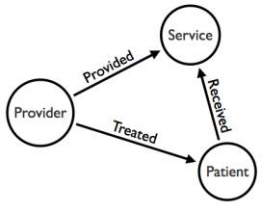
Data available:

Transactional data
Payment to provider for each service (prescription) for each patient
Limited features for services, patients, and providers

Added graph features:

Similarity of providers by services
Provider communities by patients served
Pharmacy-provider-patient tight networks

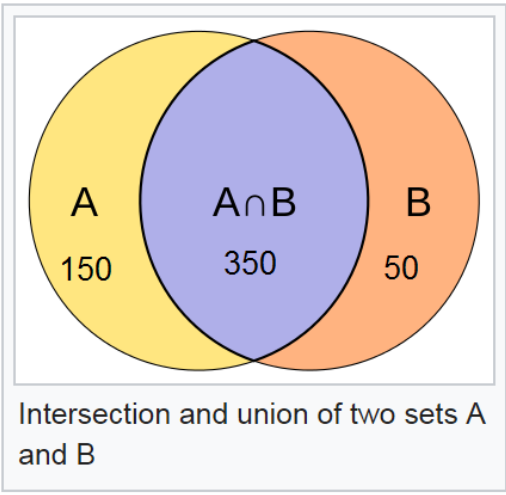
Opioid Epidemic



View the problem as a graph and investigate collusions and kickbacks

Explanation:
If you have a large amount of overlapping patients, that is anomalous

First Provider treated 500 patients
Second treated 400 patients
Of these, 350 were treated by both



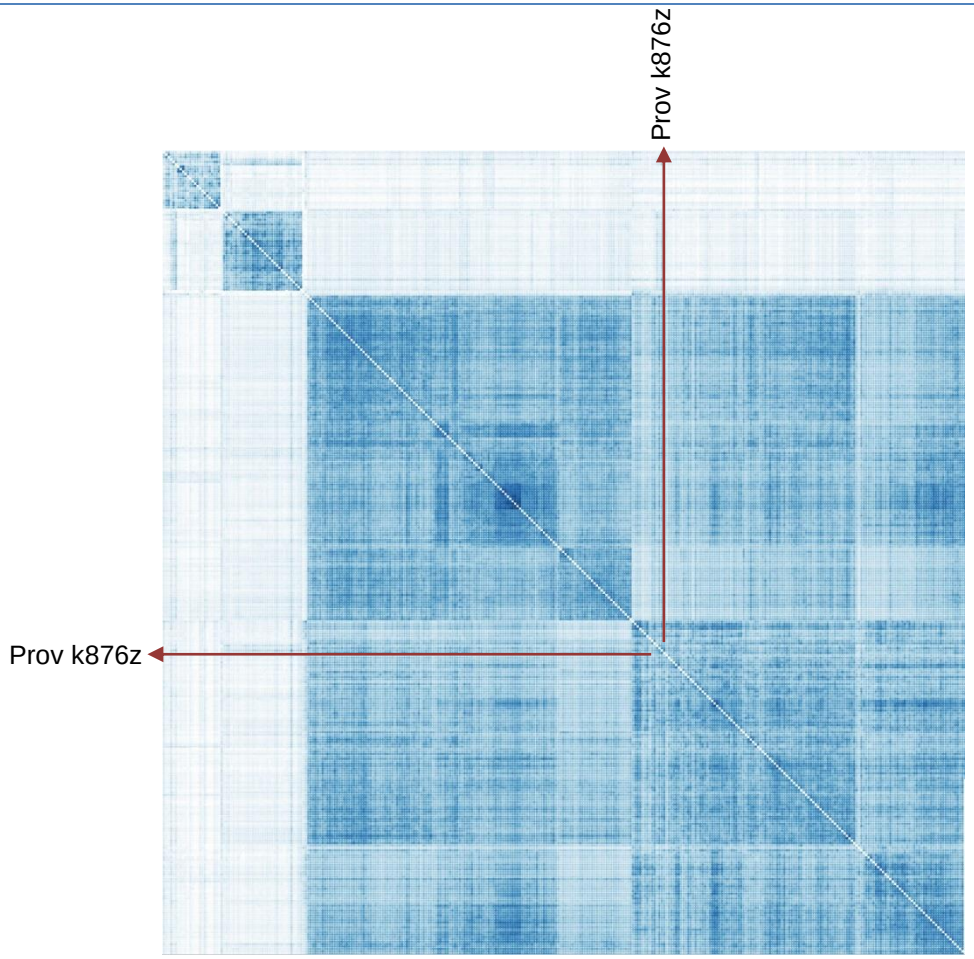
Intersection and union of two sets A and B

One simple solution: Use Jaccard Index

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

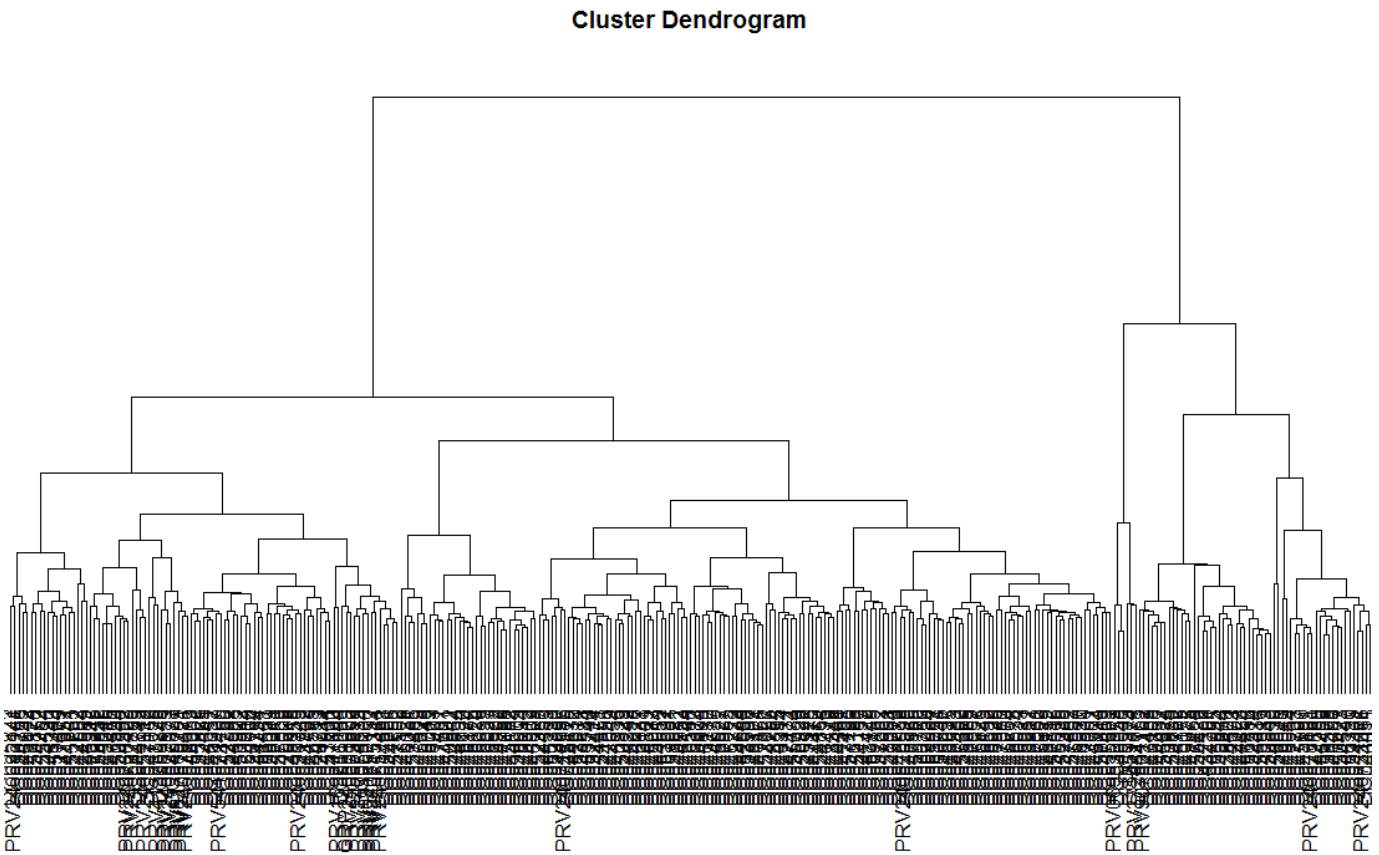
=350/(150+350+50)=0.636

Connections Reveal Communities



- Jaccard overlap of services by provider
- Sorted with hierarchical clustering
- Shows which other providers deliver similar services

Clustering by Connectedness



Putting it Together

Holistic Fraud Analytics Platform

1. Curate the data
   i. Scan internal and external sources
   ii. Gather representative samples from different times and places
   iii. Perform robust entity resolution
2. Add in relationship (graph) features
3. Remove bias from labeled data
4. Build supervised fraud model
5. Build and deploy “not good” models
6. Refresh periodically

Other Things to Consider

- Entity resolution
- Graph features
- Third party data sources
- Etc.

Q & A

Dr. Aric LaBarr, Ph.D.
Elder Research
Director and Senior Scientist
aric.labarr@elderresearch.com

Mike Thurber
Elder Research
Lead Data Scientist
mike.thurber@elderresearch.com

Blog: www.elderresearch.com/company/blog

Webinars: www.elderresearch.com/company/resource-center/webinars

Appendix
(Extra material not covered in webinar)

Fraud detection for Unemployment Insurance

Problem:

Determine who might be committing fraud for unemployment insurance

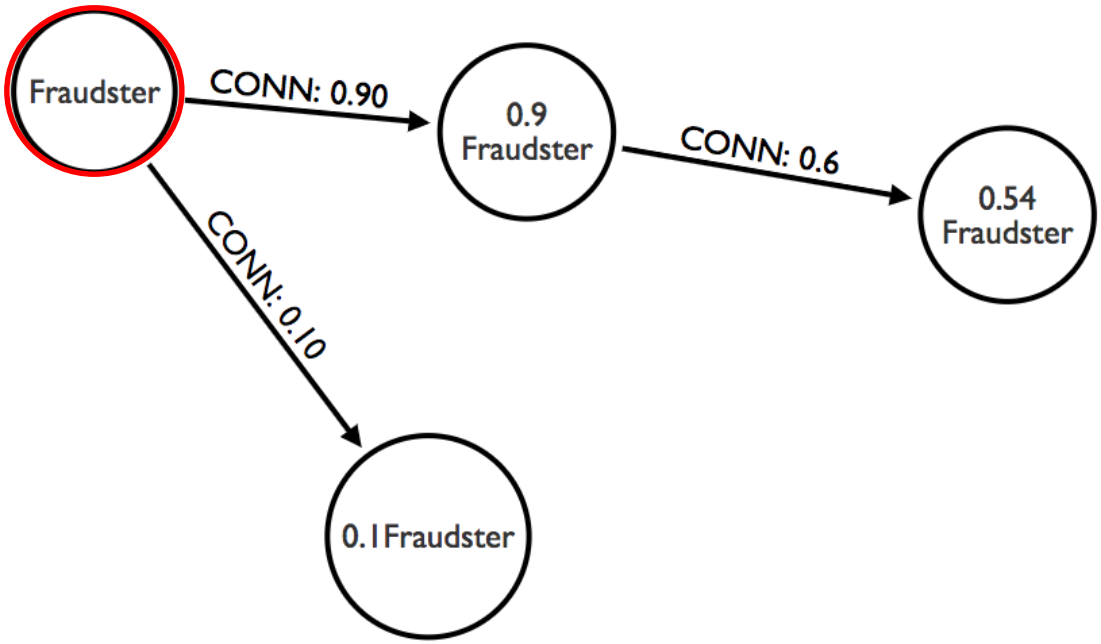
Data available:

- Had a list of known bad actors who had committed fraud (rare)
- Claimant contact information
  (phone number, email address, postal address, IP address, etc)
- Which company they worked for
- Time stamps for these records

Open Discussion with the team

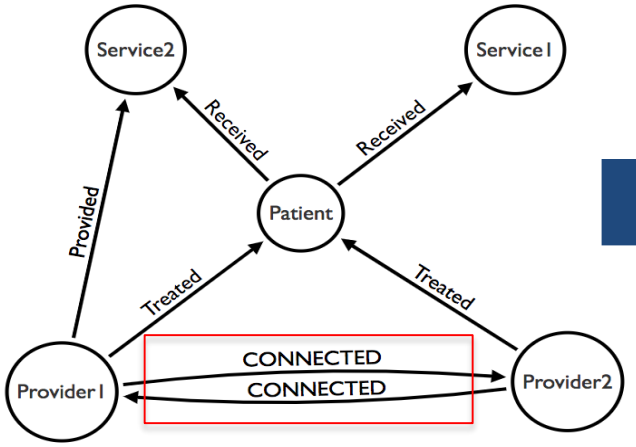
Graphs Reveal What is Connected

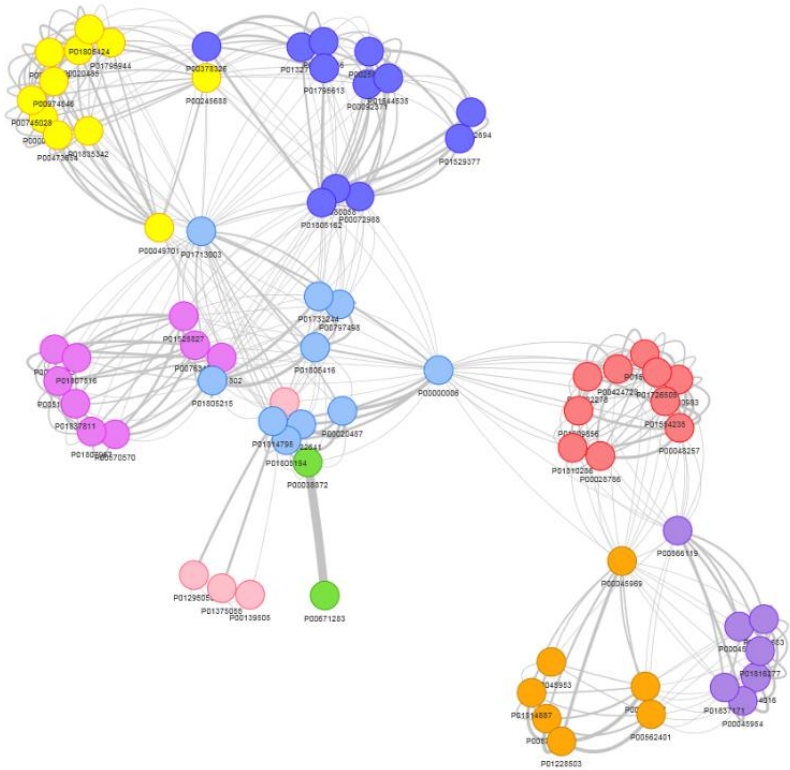
Question	Graph concept
If we know a few bad actors, who else should we monitor?	Risk propagation



Graphs Reveal What is Connected

Question	Graph concept
Which groups of people are working together?	Community detection





What Most Don't Do – But Should…

1. Build propensity to be investigated model (This is to deal with selection bias)
2. Weight "investigated claims" by 1/P[investigated]
3. Extract weighted bootstrapped sample of investigated claims
   - To be representative of all cases
4. Build fraud likelihood model using bootstrapped sample (Full model build!)
5. Explain scores (with ESP or Lime)
6. Investigate the cases most likely to be fraudulent
7. Report and label the investigated cases

Finding “Abnormal” Fraudulent Claims

1. Select labeled known good (“not fraud”) sample
   - Fraud label=“good”
2. Apply CADE, synthesizing fake records with:
   - Uniform random distribution and
   - Shuffling every column independently
   - Fraud label=“not good”
3. Build a “not known good” model
4. Check that it finds the known fraud cases
5. Explain scores (with ESP or Lime)
6. Refresh periodically

Tax Fraud

Problem:

Earned Income Tax Credit is a typical tax credit that gets abused. IRS estimates the loss is about 20 billion dollars every year. EITC is granted to to low income filers who have dependents and is a politically hot subject. Fraud is usually organized by tax preparers but IRS is having a hard time finding them as most don't trigger flags

Data available:

- Tax Preparer registration data
- Individual Tax Returns prepared by tax preparers
- Returns have information whether they filed for EITC, plus preparers' information

Tax Fraud

Our solution:

- Resolve Entities (several months of work)
- Create a network of preparers using phone, IP, EIN and addresses used
  on tax returns
- Assign weights to each relationship
- Detect communities of preparers using the graph created
- Aggregate risk scores of each preparer to the community they belong to
- Rank communities of preparers by risk score that is calculated from
  the tax returns they prepared