

CONFIDENCE IN ANALYTICS

IDENTITY DISCOVERY CHALLENGE

Final
Presentation
4/29/2013



AGENDA

- Introduction
- Overview
- Methodology
- Gephi Filter
- Validation Data
- Next Steps
- Questions

TEAM **C**ONFIDENCE **I**N **A**NALYTICS

Kate Davies

Lisa Kuhn

Betsy Matthews

John Papazian

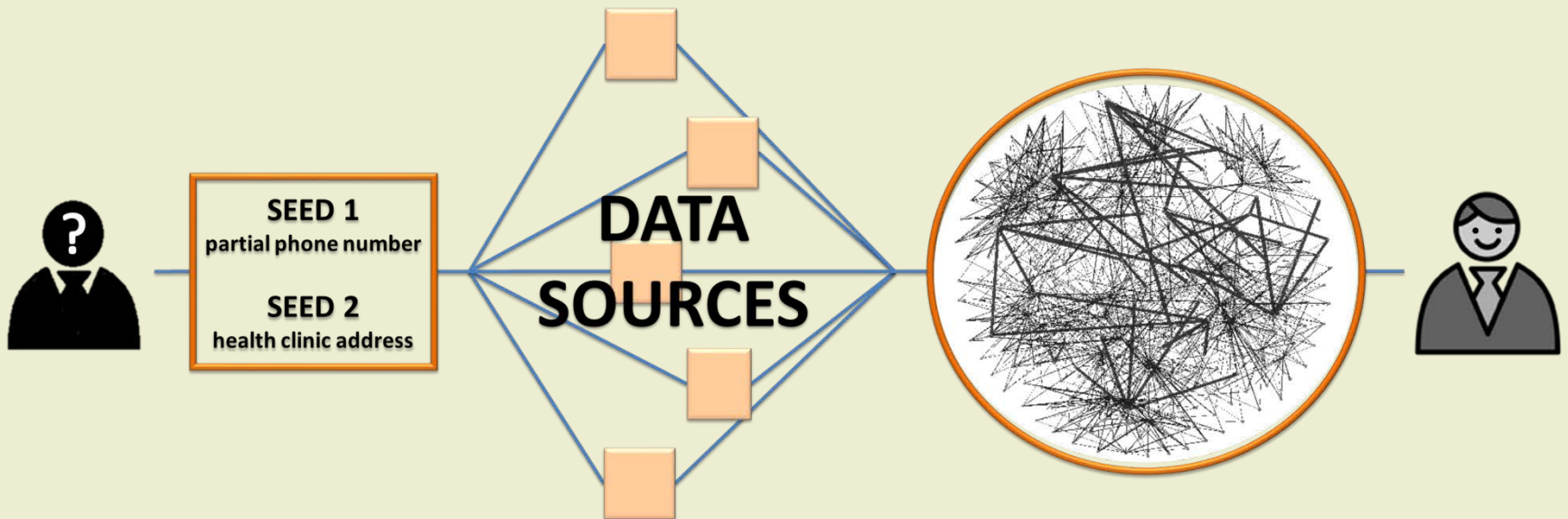
Matt Pledger

OVERVIEW

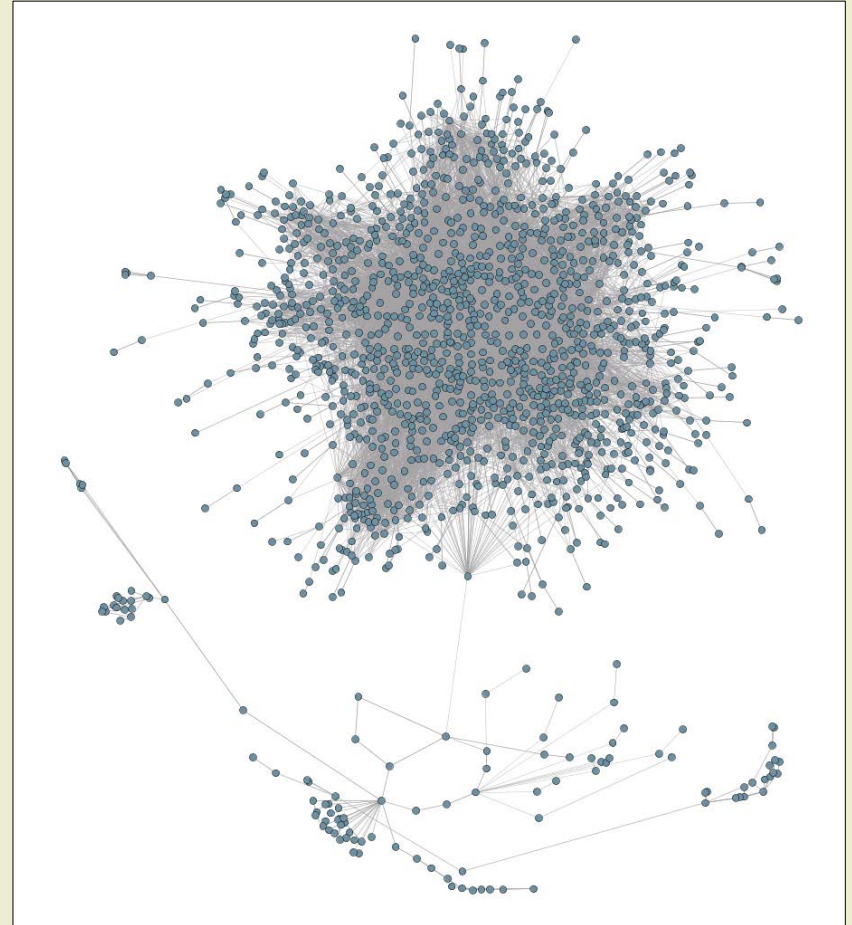
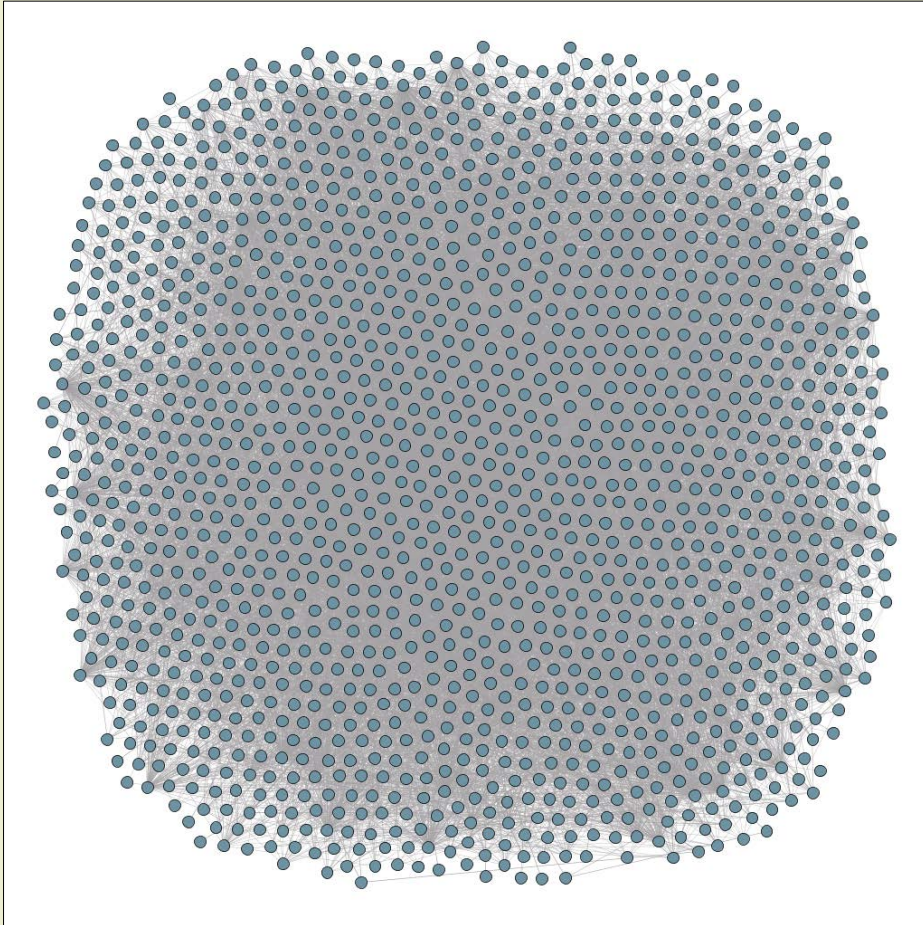
TASK

- **Alert:** Unidentified male potentially carrying a deadly and highly contagious virus
 - **SEED 1:** Partial phone number
 - (212) 998-75XX
 - **SEED 2:** Address of clinic
 - 4408 East Madison Ave, Bethesda, MD 20014
- **Task:** Identify and locate the unidentified male
 - Develop a replicable methodology
 - Find ways to visualize the identification process

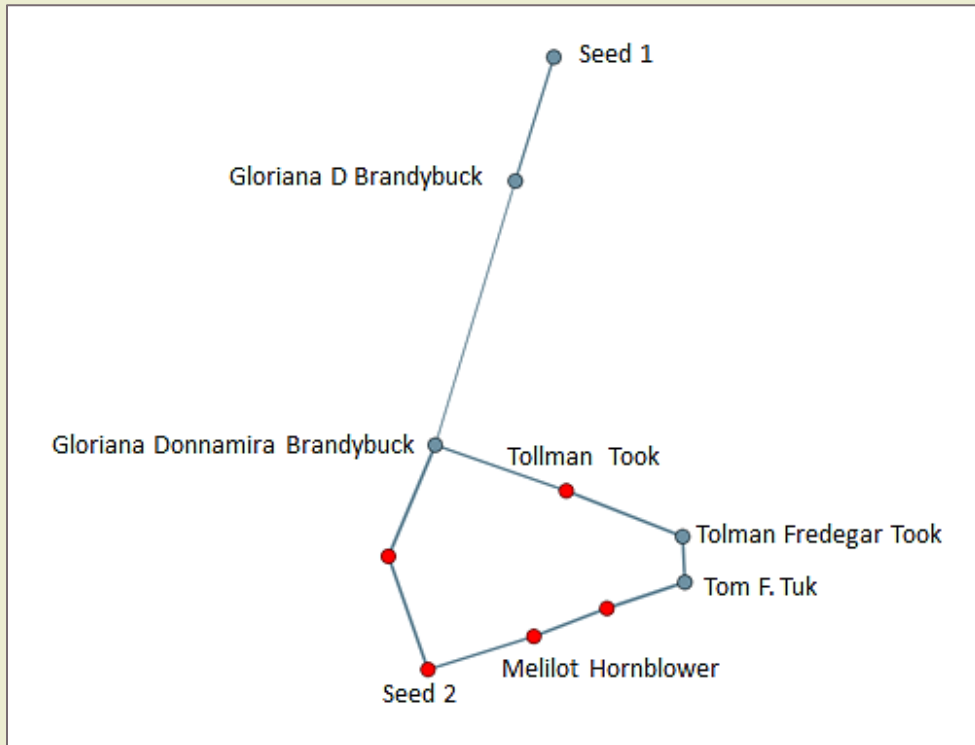
OVERVIEW ACTION



OVERVIEW ACTION



OVERVIEW RESULT



Person of Interest:
Tolman F. Took

Current Location:
322 Doe Meadow Drive
Bethesda, MD

OVERVIEW VALIDATION

- **New challenge:**

- Included 5 possible solutions
- Doubled complexity

- **Results:**

- Identified important nodes of all 5 solutions
- Less than 10 minutes using our methodology

METHODOLOGY

Data
Exploration

First Look

Dijkstra
Transversable
Subset

Anomaly
Detection

METHODOLOGY

DATA EXPLORATION

Given Data Sources



Phone Subscriber Look-Up



Credit Card Record



Credit Card Transaction



Home Purchase Agreement



Identification Document



Travel Record



Hotel Reservation

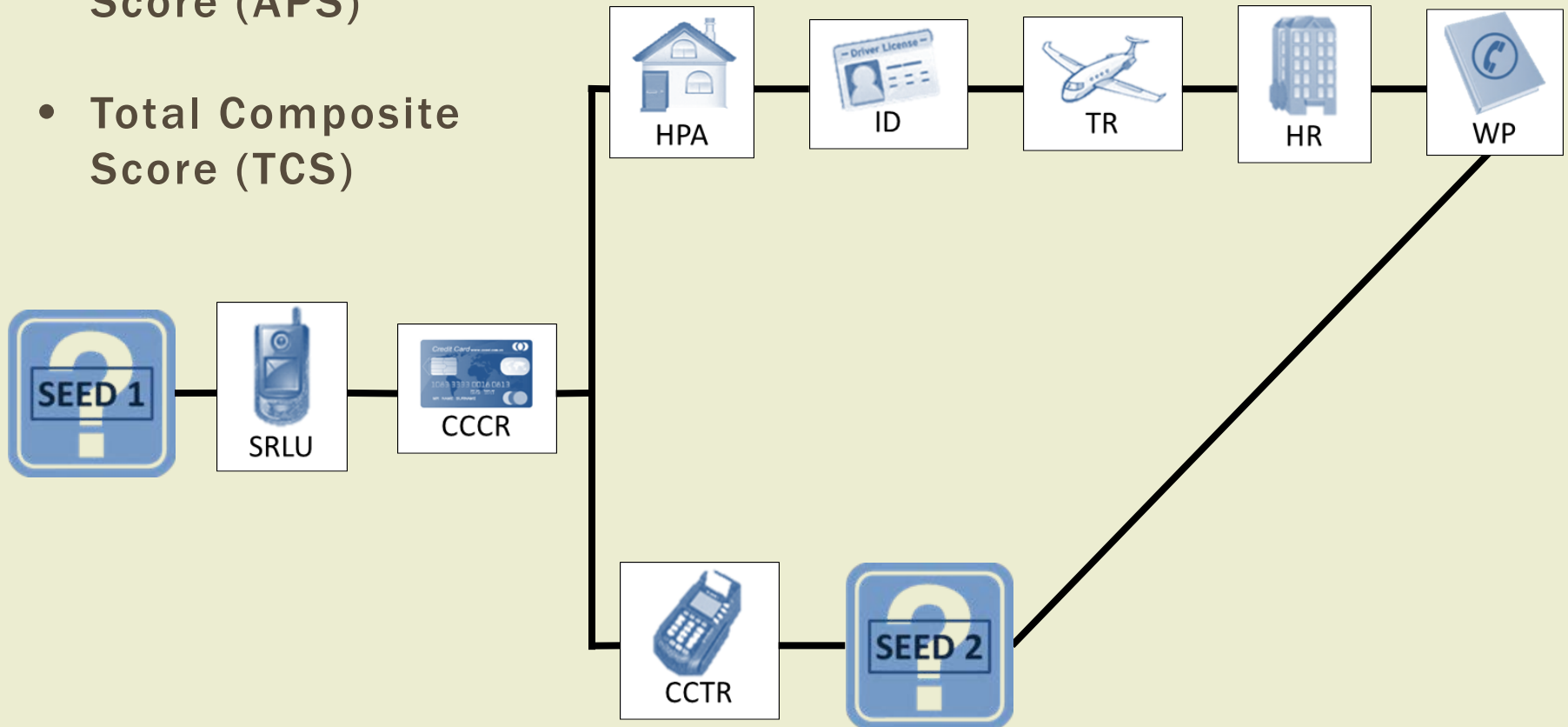


White Pages

METHODOLOGY

DATA EXPLORATION

- Attribute Pair Score (APS)
- Total Composite Score (TCS)



METHODOLOGY

DATA EXPLORATION

Shortest Path from SEED 1 to SEED 2

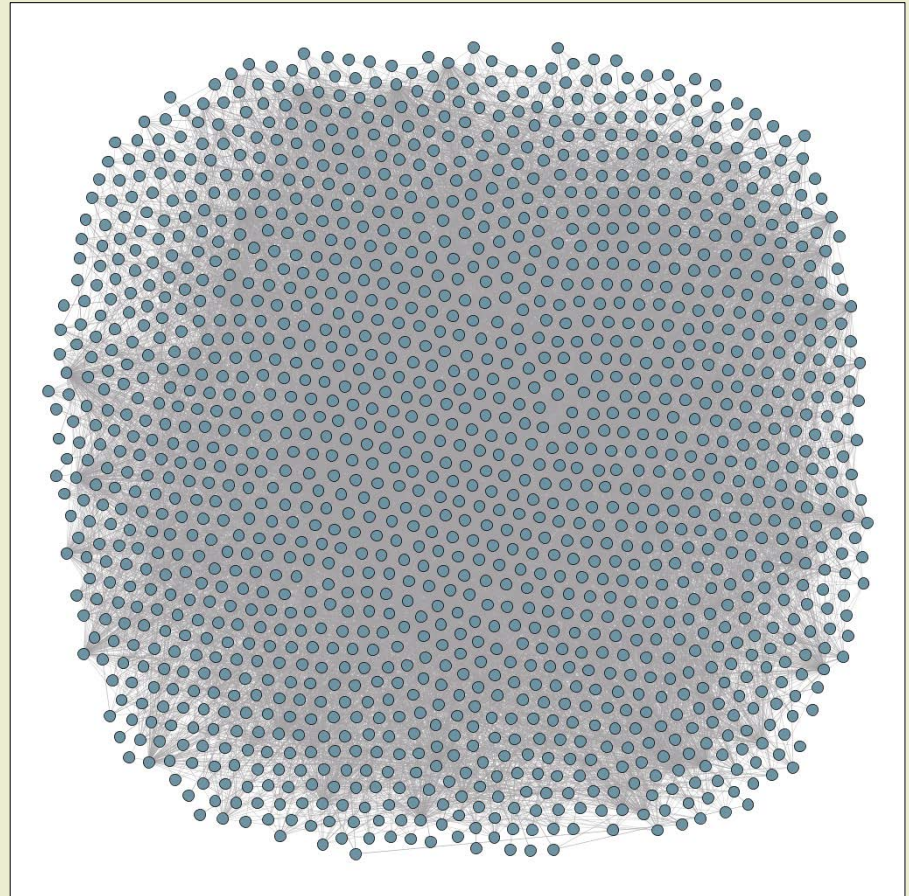
Node	Source	Link Distance	TCS	First Name	Middle Name	Last Name	Street	City	State	Zip	Phone	ID Document
1	SEED-1										21299875XX	
2	SRLU	7.20	0.71	Gloriana	D	Brandybuck	3306 Rosewood Lane	New York	NY	10003	2129987506	
3	CCCR	36.61	0.48	Gloriana	Donnamira	Brandybuck	2719 Pin Oak	Manhattan	NY	10018		5334856597493120
4	CCTR	1	1				18 Wayback Road	Bethesda	MD	20014		5334856597493120
5	SEED-2	1.39	0.95				4408 East Madison Ave.	Bethesda	MD	20014		
		46.2										

Total Distance of 46.2

METHODOLOGY

FIRST LOOK

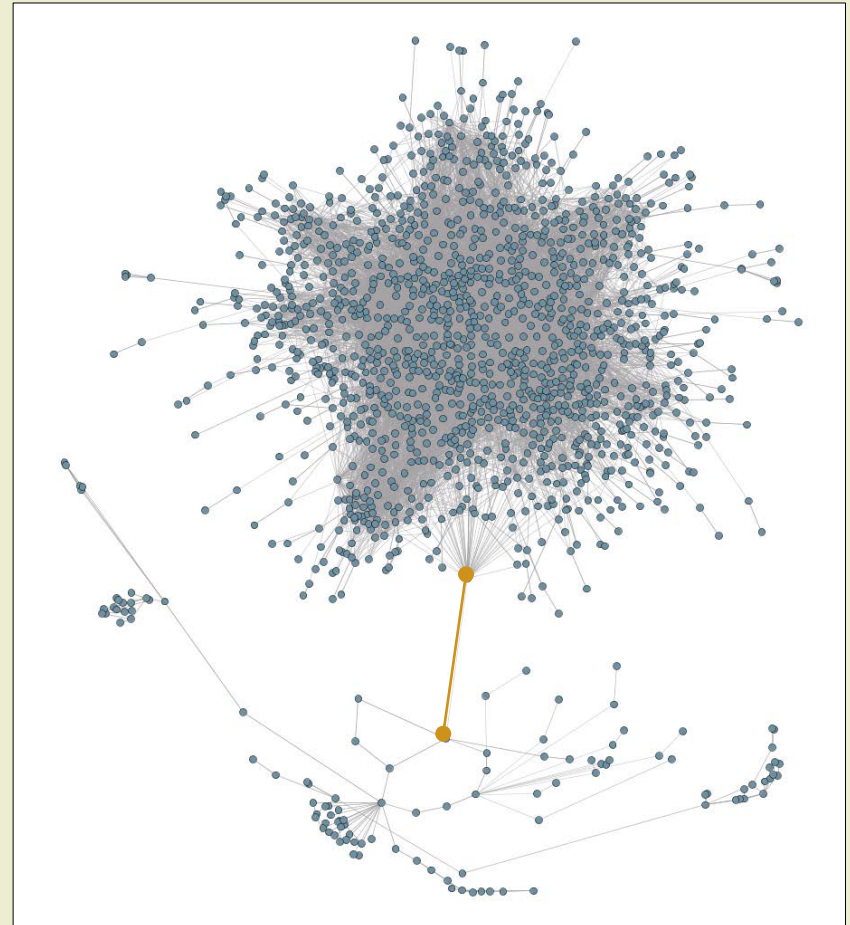
- Gephi displays our nodes and edges
- The data set contains connections from entity resolution from the 8 sources
- About 350,000 Nodes and 60,000 Edges



METHODOLOGY

DIJKSTRA TRANSVERSABLE SUBSET

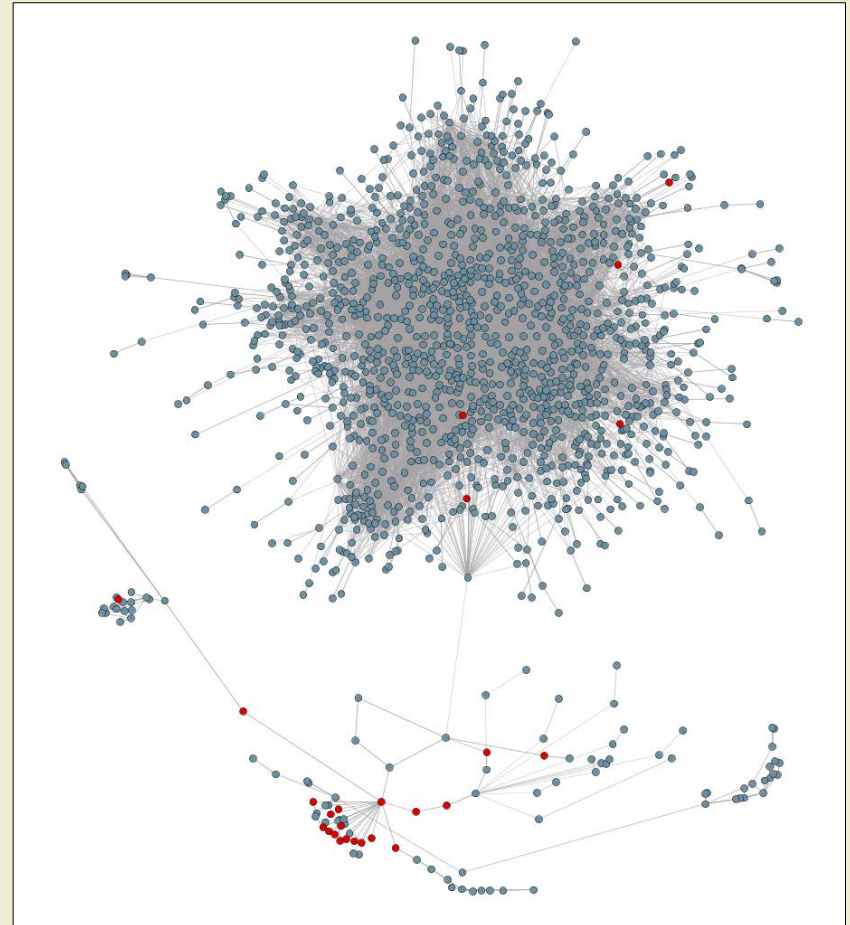
- Dijkstra's Algorithm was used to find all nodes that can be reached by the SEED nodes
- The resulting graph includes about 1,500 nodes



METHODOLOGY

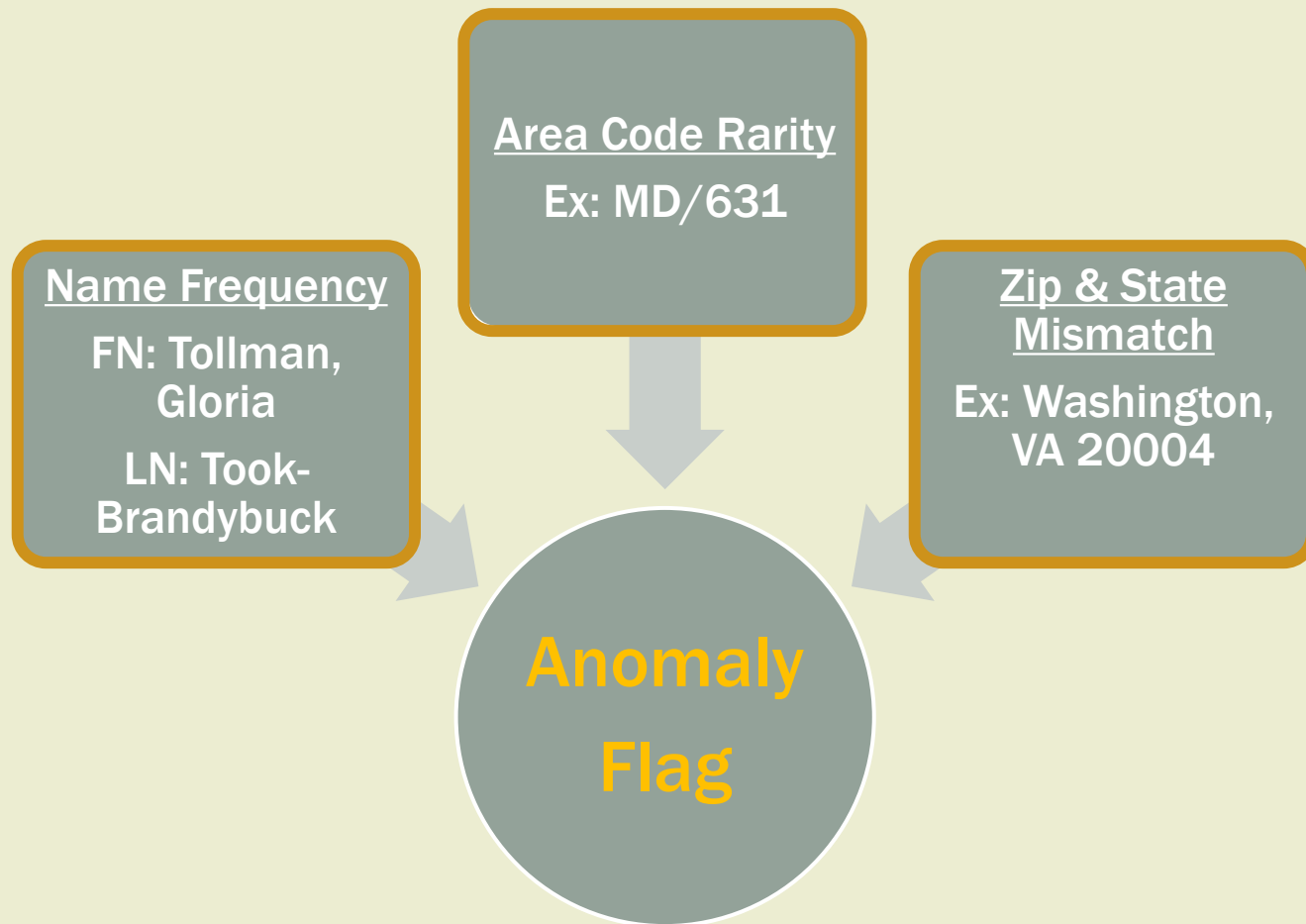
ANOMALY DETECTION

- Using SAS, we found and flagged conspicuous nodes with the following qualities:
 - Unique first name or last name
 - Unique area code within a state
 - Mismatched zip code and state



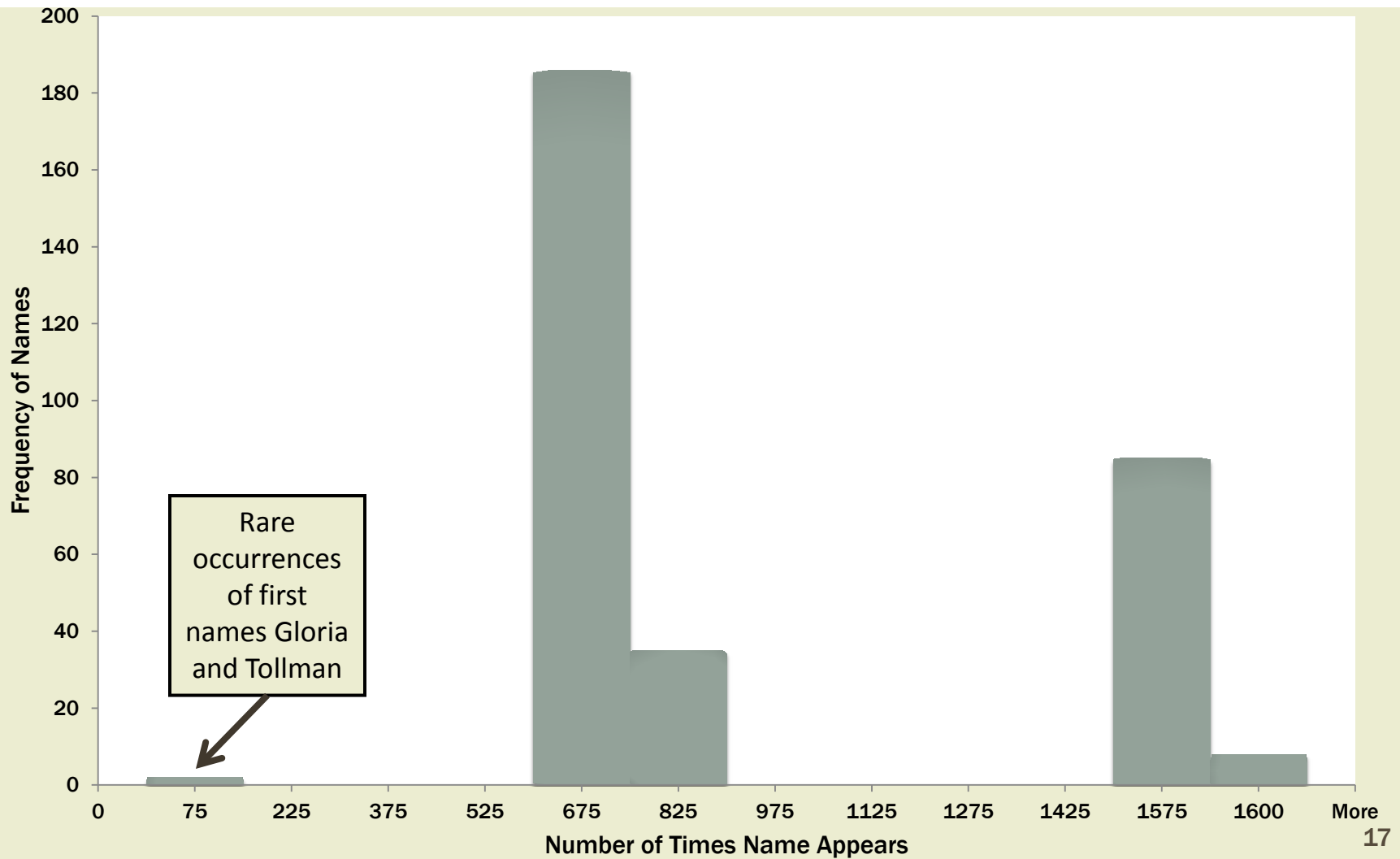
METHODOLOGY

ANOMALY DETECTION



METHODOLOGY

ANOMALY DETECTION



OUR GEPHI FILTER

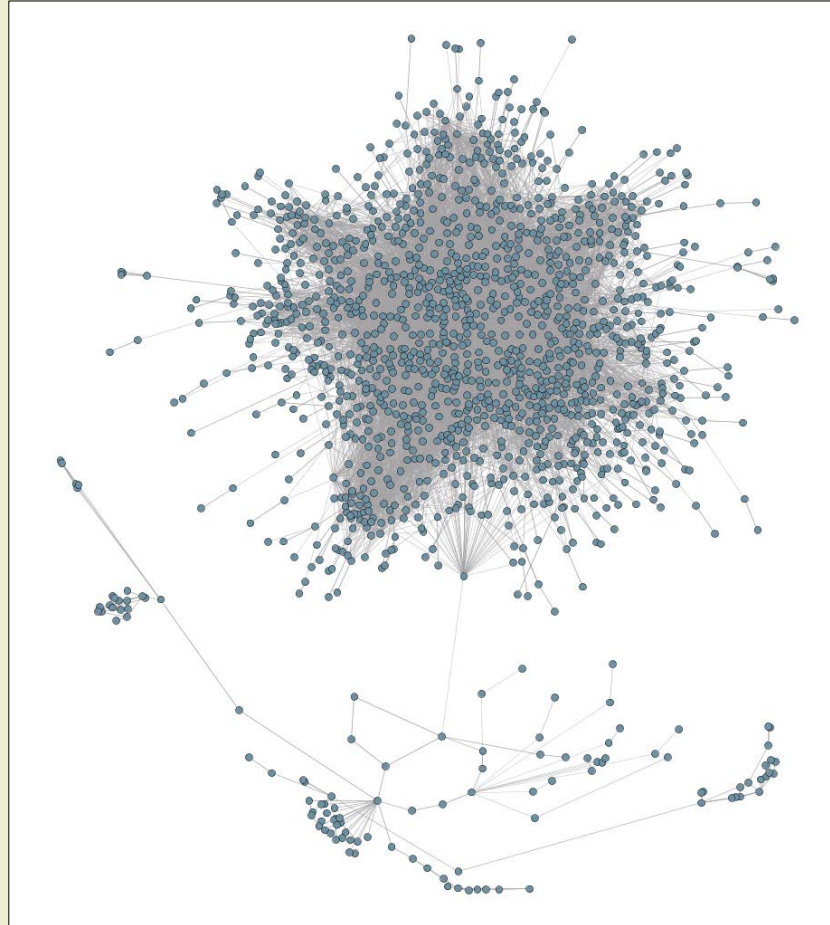
[Demo](#)

[Solution](#)

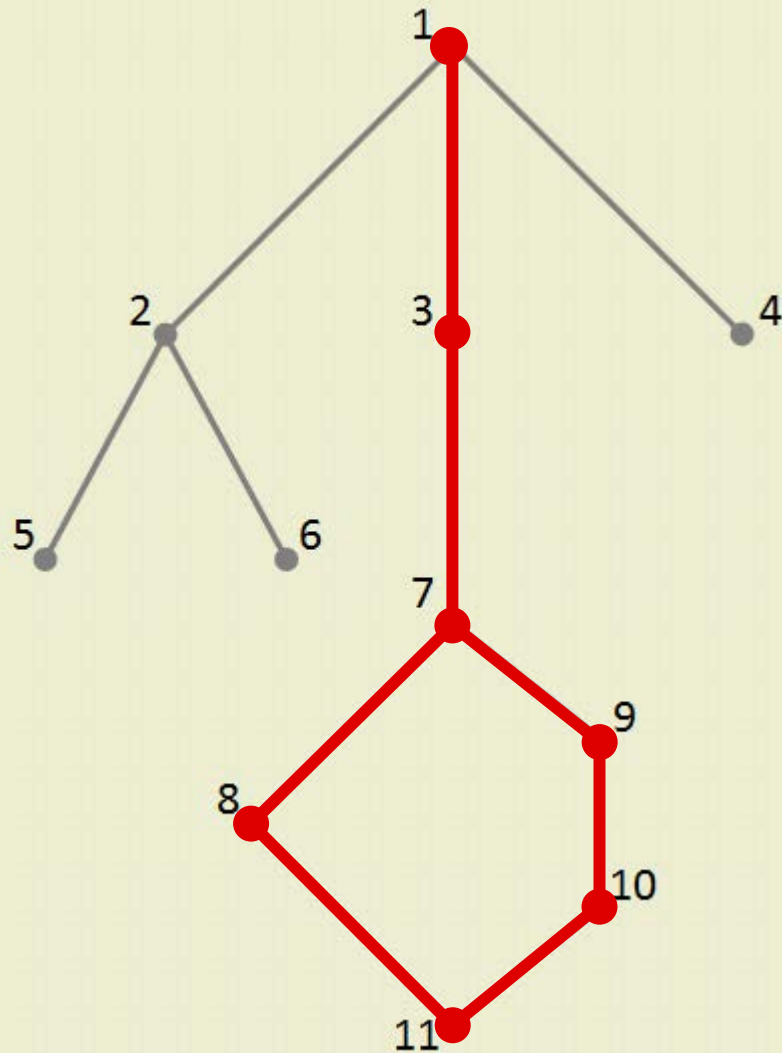
[Storyboard](#)

[Additional
Capabilities](#)

OUR GEPHI FILTER INITIAL SUBSET



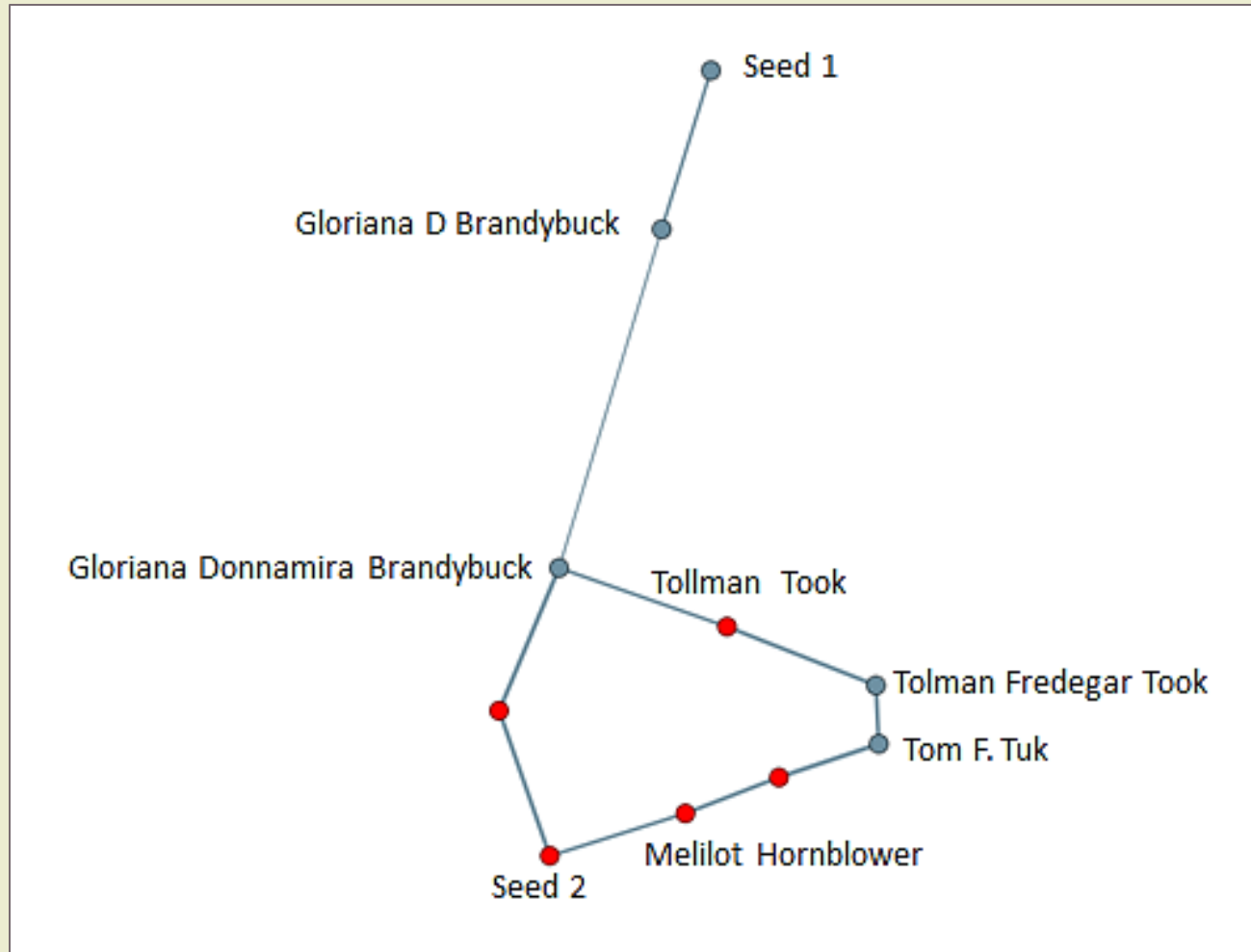
DEMONSTRATION



All Complete Paths:

$\{\{1,3,7,8,11\},\{1,3,7,9,10,11\}\}$

OUR GEPHI FILTER SOLUTION



Gloriana Donnamura Brandybuck
2719 Pin Oak Drive
Manhattan, NY 10018
5334856597493120



Gloriana D. Brandybuck
3306 Rosewood Lane
New York, NY 10003
212-998-7506



SEED-1
212-998-75XX



Gloria Took-Brandybuck
Pin Oak Dr
Manhattan, NY 10018



18 Wayback Road
Bethesda, MD 20014
5334856597493120



Tollman Took
Pin Oak Dr
Manhattan, NY 10018

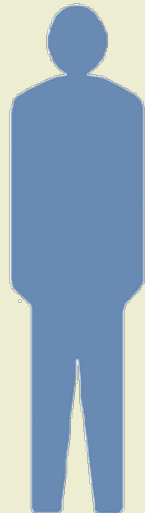


SEED-2
4408 East Madison Ave.
Bethesda, MD 20014



Person of Interest:
Tolman F. Took

Current Location:
322 Doe Meadow Drive
Bethesda, MD



Tolman Fredegar Took
234 Trails End Rd.
Staten Island, NY 10301
298808448



Melilot Hornblower
322 Doe Meadow Drive
Bethesda, MD 20014
301-803-5414



322 Meadow Dr.
Bethesda, MD 20014
631-808-5343



Tom F. Tuk
631-808-5343
298808448



OUR GEPHI FILTER ADDITIONAL CAPABILITIES

Show neighbors of important nodes

AllFullPaths Settings

Source:

Target:

Conspicuous Only

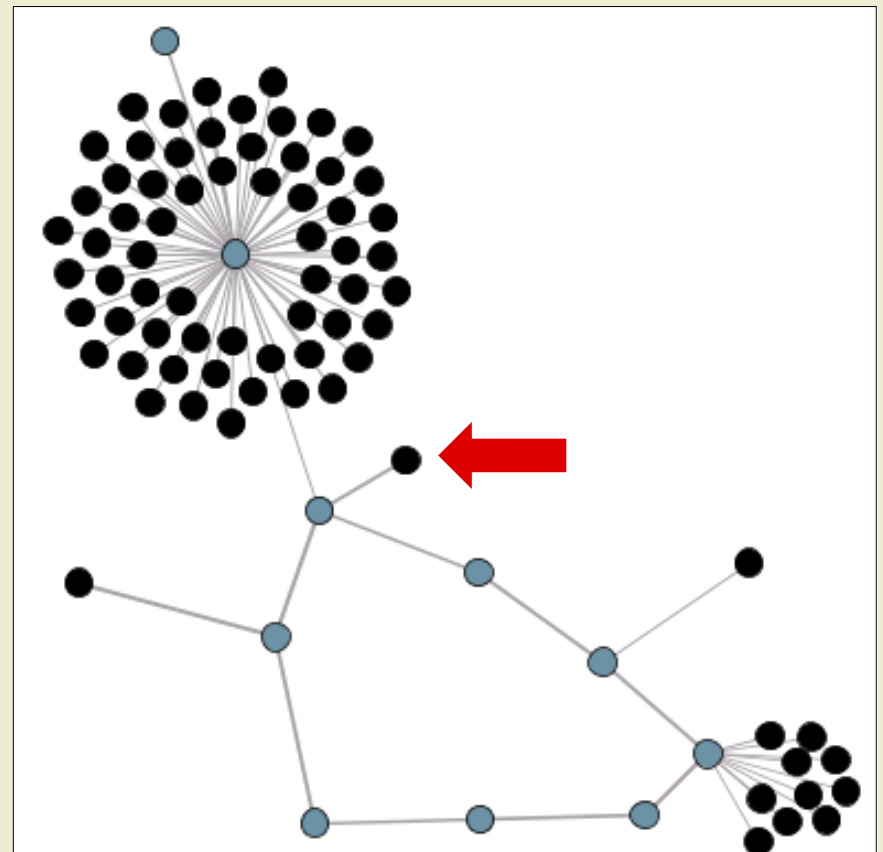
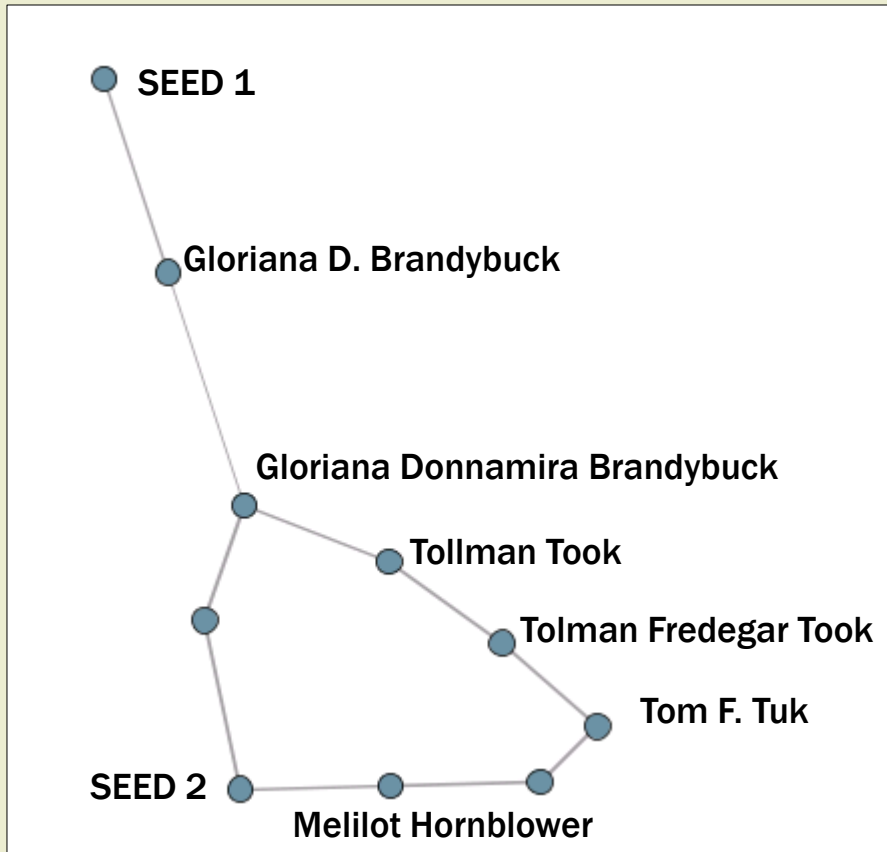
Show Neighbors

Maximum Duplicate Database Types

OK

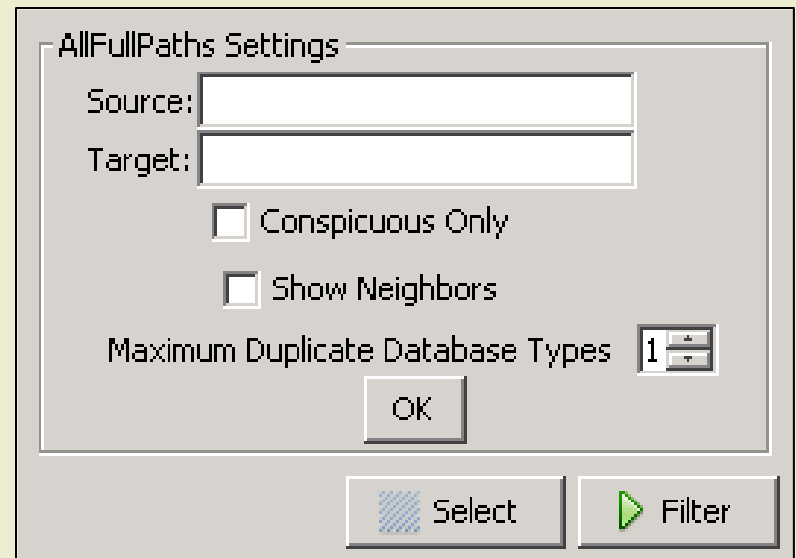
Select Filter

OUR GEPHI FILTER ADDITIONAL CAPABILITIES



OUR GEPHI FILTER ADDITIONAL CAPABILITIES

- Highlight paths with conspicuous elements
- Consider paths that have more than one node from the same data source



The screenshot shows a dialog box titled "AllFullPaths Settings". It contains the following elements:

- Two text input fields labeled "Source:" and "Target:".
- Two checkboxes: "Conspicuous Only" and "Show Neighbors", both of which are currently unchecked.
- A numeric spinner control labeled "Maximum Duplicate Database Types" with the value "1" displayed.
- An "OK" button.
- At the bottom, two buttons: "Select" (with a blue hatched icon) and "Filter" (with a green play icon).

VALIDATION DATA

Data Creation

Characteristics

Results

VALIDATION DATA DATA CREATION

Download
randomly
generated
data

Manipulate
data and
embed
solutions

Fuzzy
match data

Ensure that
embedded
solutions
exist in
Edge table

VALIDATION DATA DATA CREATION

- FakeNameGenerator.com
- KNIME with Pervasive Data Rush
- SAS (data manipulation)

FAKE NAME GENERATOR™

Step 3 - Choose name sets, countries, gender, and age

Name set

- Hispanic
- Hobbit**
- Hungarian
- Icelandic
- Igbo

Country

- Spain
- Sweden
- Switzerland
- United Kingdom
- United States**

Gender

Male: **50%** Female: **50%**

Age

19 - 85 years old

VALIDATION DATA DATA CREATION

- FakeNameGenerator.com
- KNIME with Pervasive Data Rush
- SAS (data manipulation)

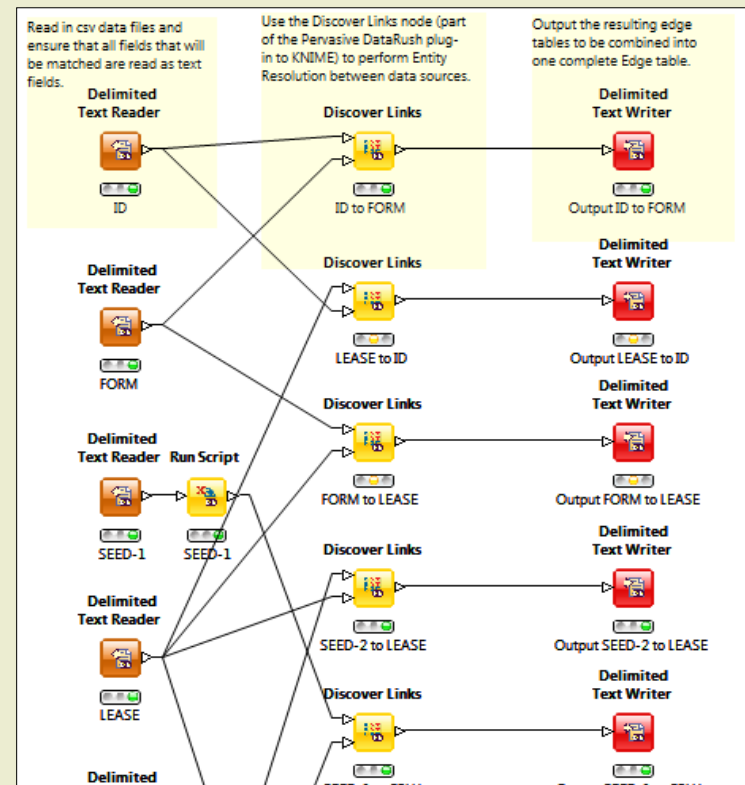


Field Comparisons

Left Input Field	Right Input Field	Comparison	Output Field	Weight
FirstName	FirstName	Levenshtein	FirstName_APS	3
MiddleName	MiddleName	Levenshtein	MiddleName_APS	2
LastName	LastName	Levenshtein	LastName_APS	3
Street	Street	Contains	Street_APS	1
City	City	Contains	City_APS	1
State	State	Contains	State_APS	1
Zip	Zip	Contains	Zip_APS	1
ID_Doc	ID_Doc	Contains	ID_Doc_APS	4
Phone	Phone	Contains	Phone_APS	2

Buttons: Add ... Edit ... Remove

Filter Value: 0.4



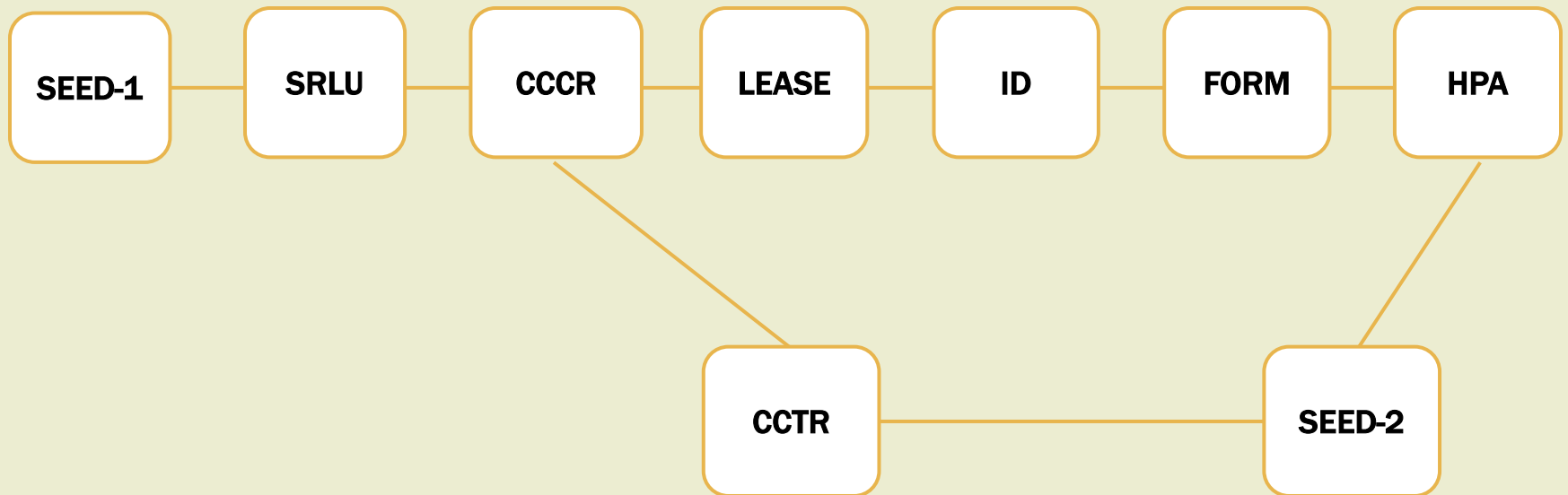
VALIDATION DATA CHARACTERISTICS

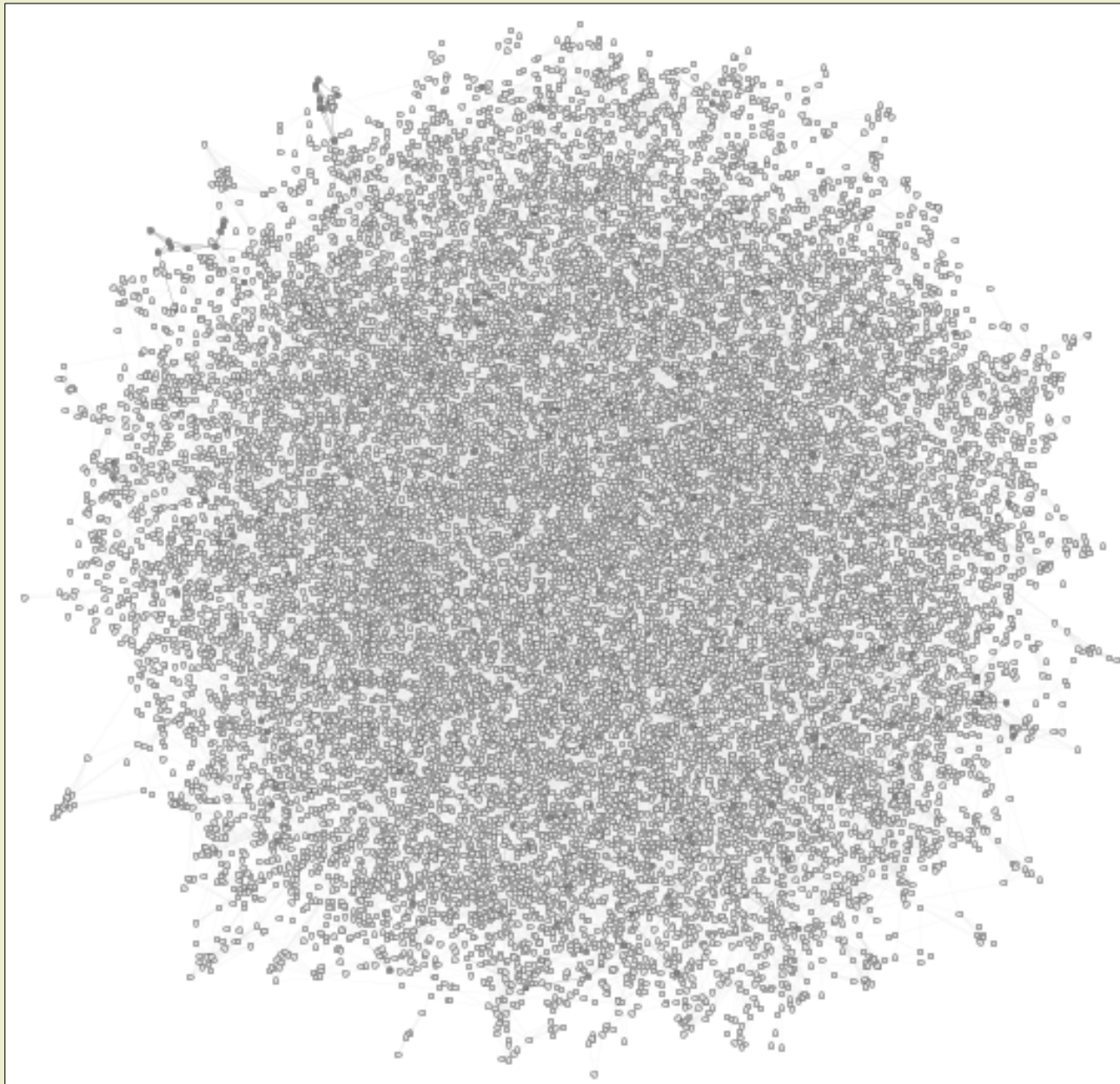
- Differences between the original and validation data:
 - Increased complexity (graph density)
 - Multiple solutions

Original	Validation
350,000 nodes 60,000 edges 1,500 node subset	300,000 nodes 38,500 edges 14,000 node subset
Edge graph density = 2.73 E -5	Edge graph density = 5.37 E -5
8 data sources	7 data sources
1 embedded solution	5 embedded solutions

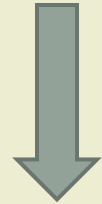
VALIDATION DATA CHARACTERISTICS

Data Source Connectivity



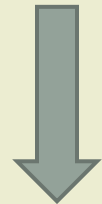


300,000 nodes

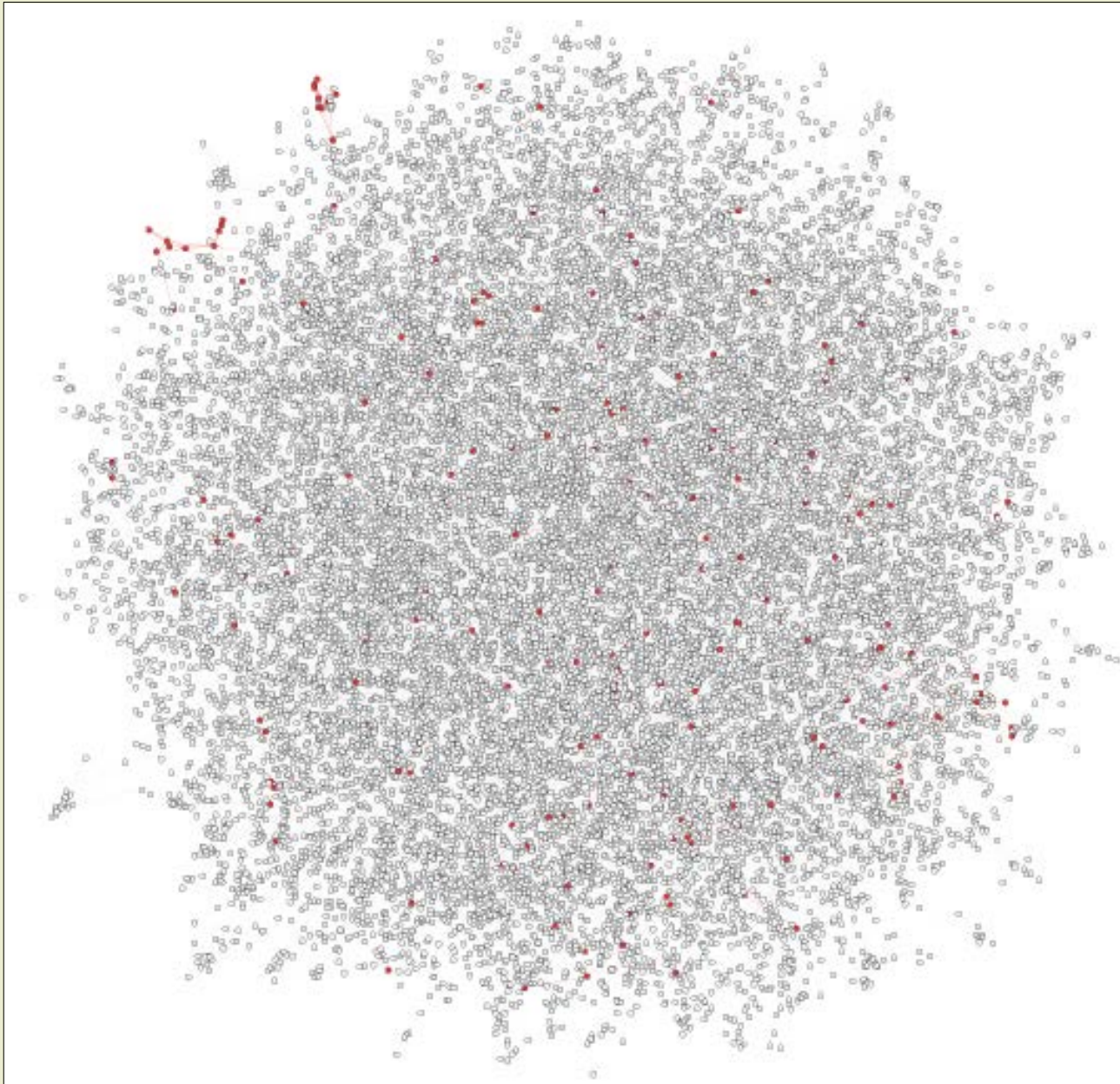


14,000 nodes

38,500 edges

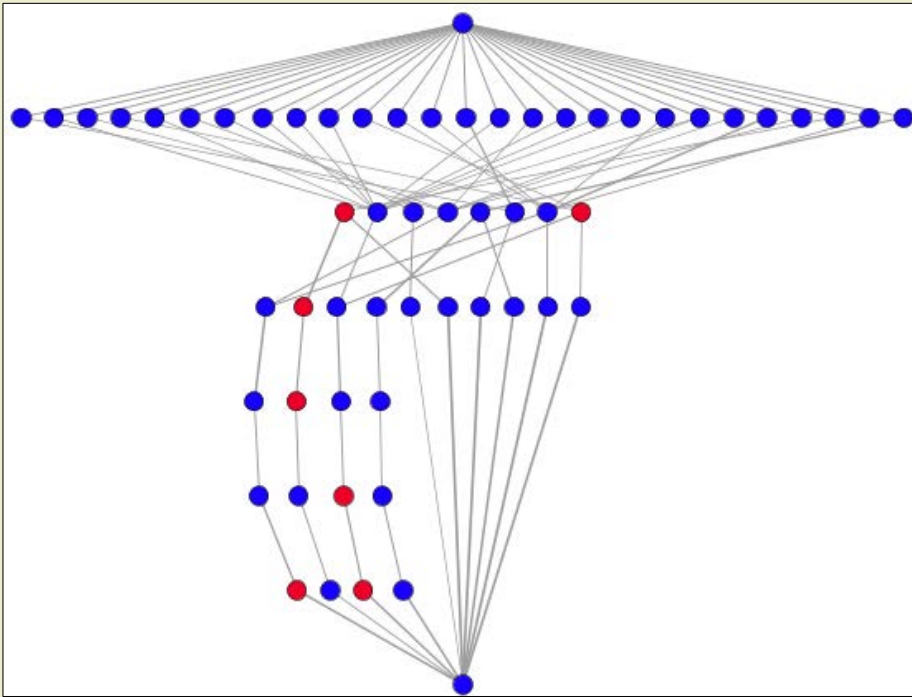


22,800 edges



0.76%
conspicuous

VALIDATION DATA RESULTS



AllFullPaths Settings

Source:

Target:

Conspicuous Only

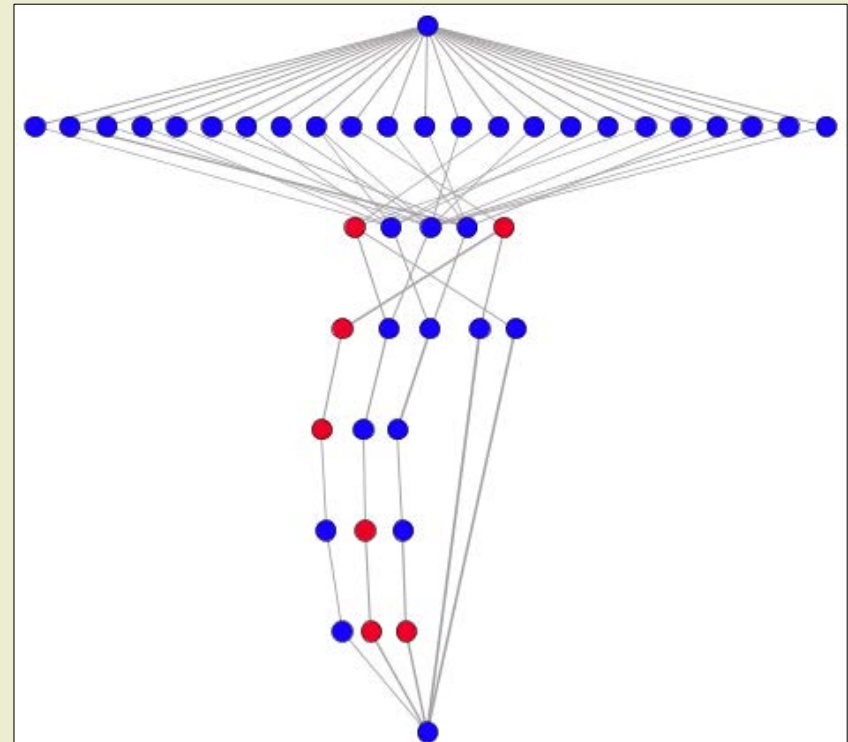
Show Neighbors

Maximum Duplicate Database Types

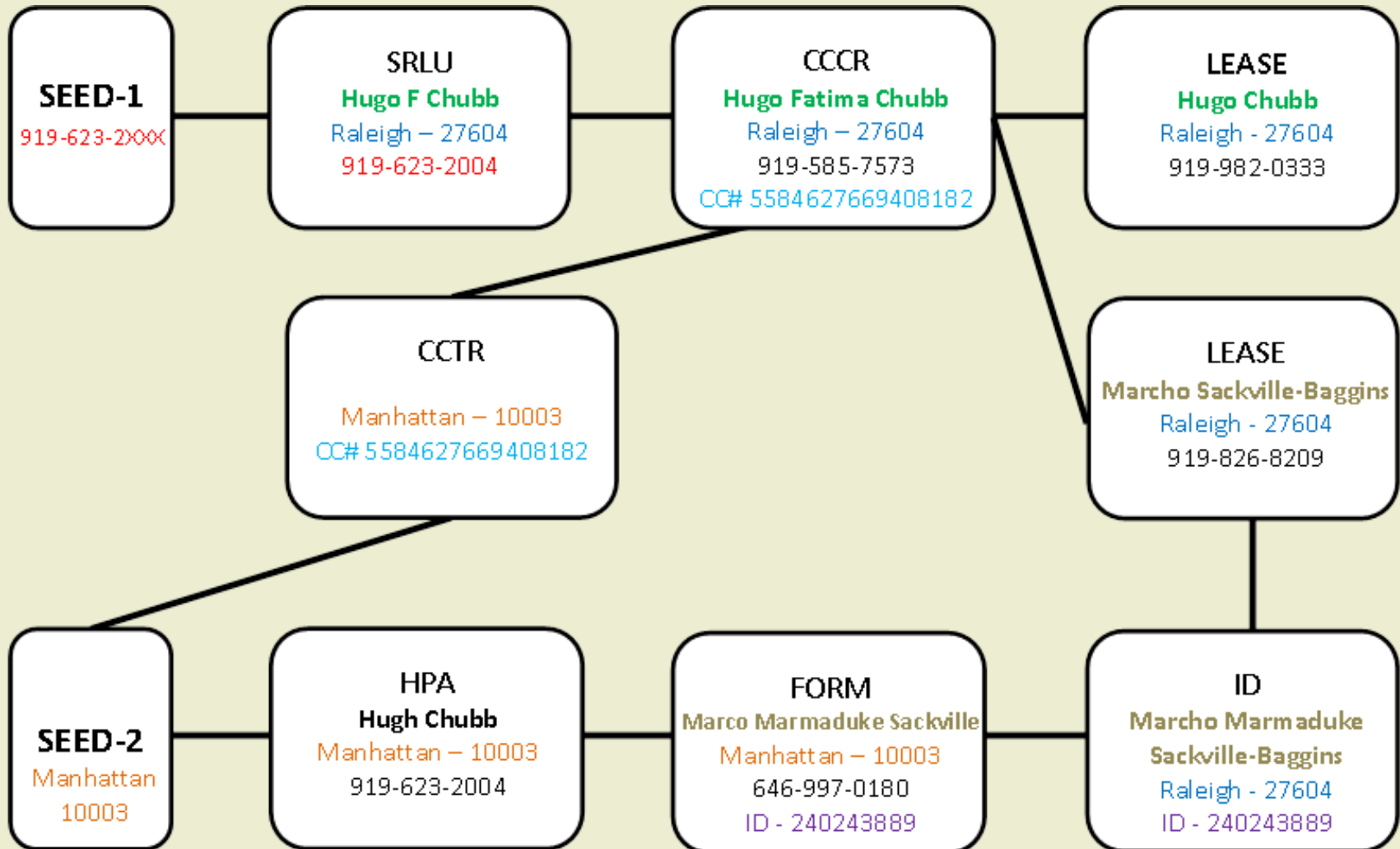
OK

Select Filter

Path reduction based on
predetermined anomaly
identification

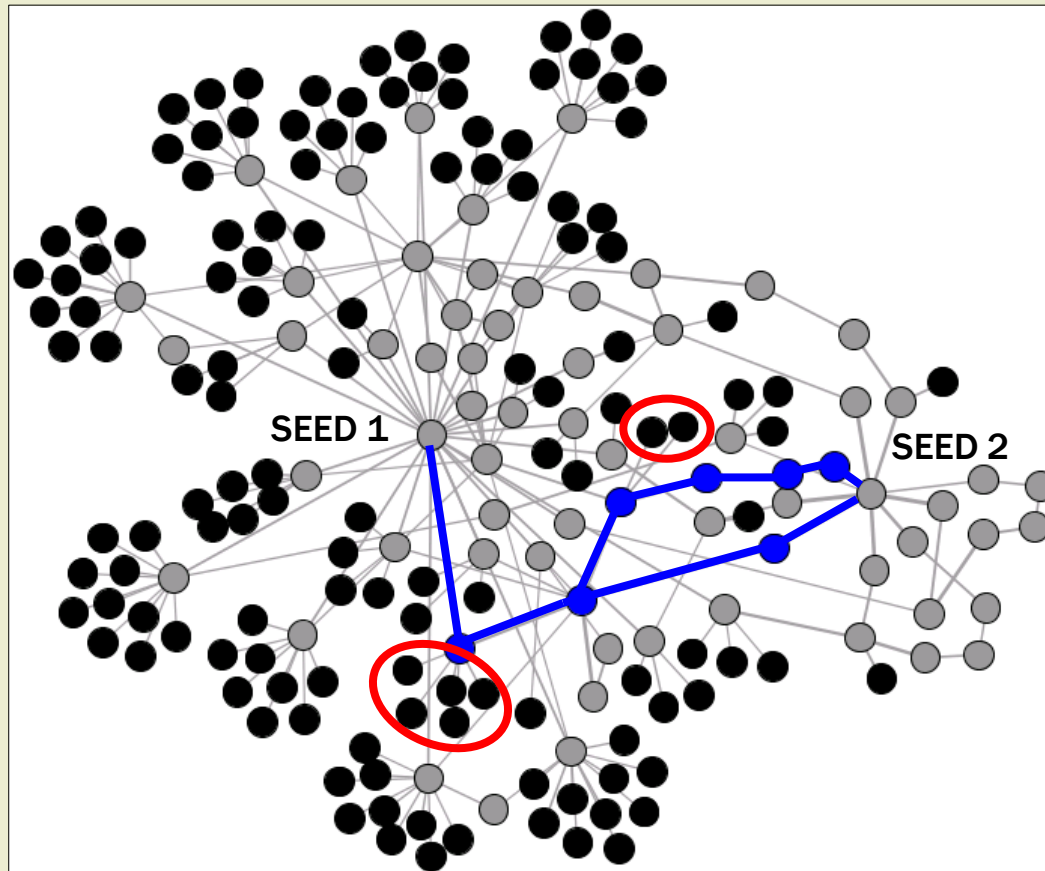


VALIDATION DATA RESULTS



VALIDATION DATA RESULTS

Show neighbors of important nodes



NEXT STEPS

NEXT STEPS

- Develop method to rank possible solutions
- Explore neighbors of important nodes further
- Find automated way to assert linkages
- Increase complexity of the data
 - Interconnectivity
 - Messy Data
 - Time Component

QUESTIONS

