The Value and Challenges of Large Scale Entity Analysis for National Security

*Perspectives Across LexisNexis*

Dr. Flavio Villanustre, LexisNexis Risk Solutions/HPCC Systems
Mrs. Mary Galvin, LexisNexis Special Services, Inc. (LNSSI)

September 19, 2012
1. LexisNexis Risk Solutions & HPCC Systems Perspectives
   • Corporate Entity Resolution Challenges
   • Enabling Technologies: LexID and the HPCC

2. LNSSI Perspectives
   • Insight into Federal Entity Resolution Challenges
   • Bringing the Power of HPCC Systems and LexID to Clients through Development of ClearView

3. Q&A
LexisNexis Overview

LexisNexis Risk Solutions

- More than 15 years of Big Data experience
- Provides information solutions to enterprise customers
- Generates about $1.4 billion in revenue
- Has been using the HPCC Systems platform for over 10 years

HPCC Systems

- Launched in June 2011
- Open source, and enterprise-proven distributed Big Data analytics platform
- To help enterprises manage Big Data at every step in the Complete Big Data Value Chain
LexisNexis Large-Scale Entity Resolution Successes

Big Data Examples
Over 4 petabytes of content (4 thousand terabytes)
- 50 billion records
- 20,000 sources
- Several million records added daily
- 7.5 billion unique name and address combinations
- 250 million unique identities
- Over 1.1 billion unique business contacts
LexisNexis Large-Scale Entity Resolution Technology: LexID

LexID℠ is the linking technology platform available with results that help you make intelligent information connections.

LexID℠ is the ingredient behind our products that turns disparate information into meaningful insights. This technology enables customers using our products to identify, link and organize information quickly with a high degree of accuracy.

LexID is the linking technology behind our products that helps customers:

Get a More Complete Picture.
Make intelligent information connections beyond the obvious by drawing insights from both traditional and new sources of data.

Better Results, Faster.
Use the fastest technology for processing large amounts of data to help you solve cases more quickly and confidently.

Protect private information.
Keep customer SSNs and FEINs secure and enjoy peace of mind knowing you are taking steps to observe the highest levels of privacy and compliance.
The Platform Upon Which LexID is Built

Designed for Big Data, Proven with Customers

High Performance Computing Cluster Platform (HPCC)

Open Source Components

- Ingest disparate Data Source
- Manage hundreds of terabytes of data
- Parallel Processing Architecture
- Discover non-obvious links and detect hidden patterns
- Language optimized for data-intensive application
- Flexibility to organize and repurpose data

INDUSTRY SOLUTIONS

- Cyber Security
- Financial Services
- Government
- Health Care
- Insurance
- Online Reservations
- Retail
- Telecommunications
- Transportation & Logistics
- Weblog Analysis

Customer Data Integration
- Data Fusion
- Fraud Detection and Prevention
- Know Your Customer
- Master Data Management
- Weblog Analysis
HPCC Systems & LexID Benefits

• Enables entity resolution on a scale not previously available
• Offers a single architecture: two data platforms (query and refinery) and a consistent data-intensive programming language, ECL
• Enables massive joins, merges, sorts and data transformations
• Accelerates creation of new services via rapid prototyping capabilities
• Cost Savings: Commodity hardware and fewer people can do much more in less time
LNSSI Insight Into Federal Entity Resolution Challenges & How HPCC Systems and LexID Can Help
Entity Resolution Challenge #1: Precomputing Correlations vs. Correlating at Query Time

**Precomputing Correlations**

- **Pros:**
  - Reduces the order of the graph for searching
  - Is optimal when relationships are defined as lines (links between entities) and there is a need to uncover all of them

- **Cons:**
  - Not cost effective if a large % of correlations are never used

**Correlating at Query Time**

- **Pros:**
  - Well-suited for searching triangles (or greater)
  - Useful when there is only interest in a tiny portion of a large dataset

- **Cons:**
  - Considerable performance hit compared with the precompute technique
## LexisNexis Approach to Precomputing Correlations

### Step 1: Generate specificity values

**Average Field Specificity:** Accounts for the distribution of unique values for each field; the more the values within a field vary, the more influence that field will have on potential record matches.

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Field Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patronym #1</td>
<td>15.48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tariq</td>
<td>3</td>
<td>21.49</td>
</tr>
<tr>
<td>Saeed</td>
<td>6</td>
<td>20.49</td>
</tr>
<tr>
<td>Fayed</td>
<td>6</td>
<td>20.49</td>
</tr>
<tr>
<td>Hussein</td>
<td>10</td>
<td>19.75</td>
</tr>
<tr>
<td>Naser</td>
<td>18218</td>
<td>8.92</td>
</tr>
<tr>
<td>Khalid</td>
<td>77458</td>
<td>6.84</td>
</tr>
</tbody>
</table>

**Field Value Specificities:** Calculated for each unique value contained in the field across all records in the dataset; the more unique a value, the more influence that value has on potential matches.
Step 2: Define relationships

Specify the number of required links before the relationship between two clusters will be considered valid.

Set the required sum of matching thresholds before a relationship will be considered valid.

Assign field(s) to define the relationship.
Step 3: Compute Correlations & Relationships
LexisNexis Approach to Precomputing Correlations

Step 4: Analyze correlation and relationship results
Step 5: Re-tune correlation strategy, if necessary

a. Methods for Refining the Record Linking Strategy:
   - Modify thresholds
   - Enable/disable fuzzy options
     - Bag of Words
     - Propagation
     - Initial
     - Edit1/Edit2
   - Create “concepts”
   - Modify field specificity values directly

b. Methods for Refining Relationship Strategies:
   - Modify fields upon which relationships are based:
LexisNexis Approach to Performing Correlations at Query Time

Correlating triangular relationships or higher is only feasible at query time: $N + \frac{N(N-3)}{2}$

LexisNexis Knowledge Engineering Language (KEL) is well-suited for solving above
Other Entity Resolution Challenges of LNSSI Clients

- What to do when all that is known about entities are their relationships, and standard biographical information is not necessarily available

- How to more effectively apply the dimensions of space and time to the entity resolution problem
Q & A
Contact Information

Dr. Flavio Villanustre
p: 770-752-3320
flavio.villanustre@lexisnexis.com

Mary Galvin
p: 202-595-4043
mary.galvin@lnssi.com
Reference Slides
Public Data Social Graph

- **Social Graph Overview**
  - What is a Social Graph.
  - Examples of Social Graph Analytics seen every day.

- **LexisNexis Public Data Social Graph (PDSG)**
  - Relationships derived from 50Tb of Public Data.
  - High Value relationships for Mapping trusted networks.

- **Large Scale Data Fabrication and Analytics**
  - Thousands of data sources to ingest, clean, aggregate and link.
  - 300 million people, 4 billion relationships, 700 million deeds.
  - 140 billion intermediate data points when running analysis.

- **HPCC Systems from LexisNexis Risk Solutions**
  - Open Source Data Intensive high performance compute supercomputer. (http://hpccsystems.com)

- **Innovative Examples leveraging the LexisNexis PDSG**
  - Healthcare
    - Medicaid\Medicare Fraud
    - Drug Seeking Behavior
  - Financial Services
    - Mortgage Fraud
    - Anti Money Laundering
    - “Bust out” Fraud
Medical Fraud

Scenario
Proof of concept for Office of the Medicaid Inspector Generation (OMIG) of large Northeastern state. Social groups game the Medicaid system which results in fraud and improper payments.

Task
Given a large list of names and addresses, identify social clusters of Medicaid recipients living in expensive houses, driving expensive houses.

Result
Interesting recipients were identified using asset variables, revealing hundreds of high-end automobiles and properties.

Leveraging the Public Data Social Graph, large social groups of interesting recipients were identified along with links to provider networks.

The analysis identified key individuals not in the data supplied along with connections to suspicious volumes of “property flipping” potentially indicative of mortgage fraud and money laundering.