Paper 4724-2020

Enhance Customer Risk Scoring by Quantifying the Opportunity to Commit Financial Crime

Steve Overton, 6 Degree Intelligence

ABSTRACT

Analytical scoring models help organizations quantify risk and make appropriate data-driven decisions across many areas. Fraud and financial crime related scoring models consist of complex formulas and mathematical equations designed to measure an outcome or target activity using multiple inputs and dimensions. Traditional scoring models measure commonly accepted dimensions across known activity, demographics and product behavior but often do not consider the means and opportunity for a criminal to transact through other entities or networks of bad actors. This paper will present a methodology and propensity modeling approach which enhances financial crime risk scoring models by analyzing the framework and infrastructure available to commit financial crime.

INTRODUCTION

Statistical models are used to predict and analyze situations and activities across many industries. This paper focuses on processes and techniques which extend traditional models by providing advanced risk-based dimensions leveraging network analytics. Techniques described in this paper can be used to understand and monitor risk within financial crime related areas such as Bank Secrecy Act (BSA) Compliance, Customer Due Diligence (CDD), Fraud and Know Your Customer (KYC). The models and approaches presented in this paper provide possible routes to take when quantifying risk and opportunity. Ideas and concepts should be well-tested and approved before implementing.

This paper is intended for experienced SAS users designing models and tools for measuring and predicting risk within a financial crime business unit. Readers of this paper should also have a basic understanding of node-link network style data references and graph theory principles. Further documentation regarding graph data modeling and node-link network style data nomenclature is provided in the References section.

Every organization is different and has its own set of unique challenges and risk profile. The goal of this paper is to provide a blueprint for success by sharing lessons learned, potential solutions and new ideas. Suggestions provided in this paper should be discussed and vetted with key stakeholders and decision makers to help navigate internal approval processes and ensure outcomes are aligned with management. Examples given in this paper are based on experiences developing financial crime systems such as Anti-Money Laundering (AML) transaction monitoring, Customer Due Diligence risk scoring and fraud detection within the banking industry.

KNOW YOUR CUSTOMER BACKGROUND

Banking and financial services organizations follow Know Your Customer (KYC) provisions of the USA PATRIOT Act to reduce the risk and possibility of the financial system being used for money laundering, terrorist financing or other financial crime related activities. KYC policies are different for each institution due to many different factors specific to the respective business and risk profile. KYC risk-based procedures include thorough onboarding steps to gather and collect customer information and review risks associated with doing business with respective customers. Customer risk is assessed, analyzed and scored through a process known as Customer Due Diligence (CDD). Common CDD risk

categories include products and services, customers and entity demographics, and geographical location. A more stringent analysis is performed for customers which present an elevated level of risk through a process known as Enhanced Due Diligence (EDD). EDD investigations add additional layers of ongoing periodic monitoring. All processes and procedures described above are heavily audited, regulated and monitored. Organizations are constantly trying to improve and streamline these processes through ongoing model risk management while maintaining budget.

FOLLOW THE MONEY FASTER

This paper presents analytical solutions to the overall model risk management challenge by providing explainable methods to monitor complex commercial risk effectively while focusing on the overall goal of preventing financial crime. Complex criminal infrastructures operate longer and better because smart criminals know how to use the financial system to their advantage by layering relationships and acting through multiple identities. As a conceptual example, Figure 1 shows complex layers of people and shell companies which can be setup and layered across many jurisdictions to distance and hide identities from monitoring systems. Monetary transactions can flow through many individuals and commercial entities that are connected across social and legal networks. Network analysis shifts the lens of the investigator or automated monitoring system by refocusing on the overall framework or opportunity available to individuals being monitored and activity across the network, not just within a certain segment or single point. For example, a person or corporate entity with hundreds of companies spread across dozens of jurisdictions presents a very different level of risk compared to someone with a few companies or none at all.

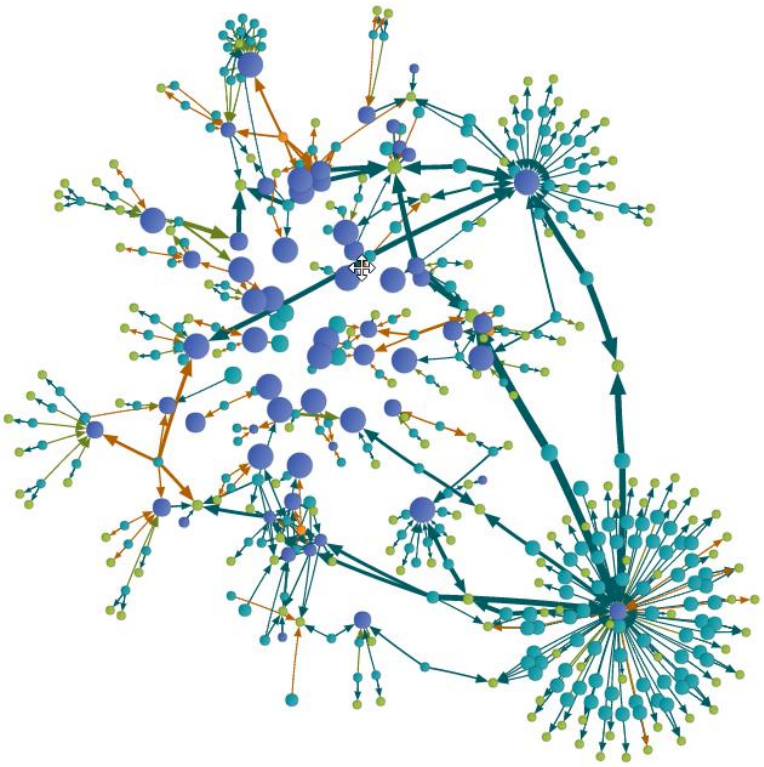

Figure 1: Example complex corporate infrastructure from the Panama Papers.

In summary, one of the primary outcomes of KYC is a predictable risk-based approach to quantify a customer's ability to commit financial crime. The corporate infrastructure or social framework available to a person or potential criminal is a critical dimension that can be used to enhance CDD processes and analytical models. The individual or single entity has been monitored for years and potentially over-monitored in some institutions. Using corporate data with advanced network analytics provides an explainable risk dimension that connects the dots across complex relationships to follow the money faster.

For additional background information on KYC, CDD and EDD processes reference the FFIEC BSA / AML Examination Manual.

EVERYTHING STARTS AND ENDS WITH THE DATA

Data is the fuel which powers the analytical engine. Sound data management practices are needed for an operationally smooth analytical lifecycle. Many aspects presented in this paper assume data is available and matured to the point it can be used to accurately model and analyze risk.

GRAPH DATABASE VERSUS RELATIONAL DATABASE CONSIDERATIONS

This paper leverages concepts surrounding network graph theory. There are many graph database storage technologies in the marketplace which are designed to store graph style data. Graph databases can be leveraged to store information for analytical purposes and can show value very rapidly assuming the analytical tools used are compatible with the graph database technology deployed or the graph database has a native querying language which satisfies analytical needs. Alternatively, traditional relational databases can be leveraged to physically store data while using logical and conceptual modeling techniques that mimic the node link structure of graph databases.

SAS® Viya does not currently support graph databases directly through traditional SAS/ACCESS® engines. Creative methods to reach graph databases can be developed in SAS through different developer API calls but may be more complicated than anticipated and suffer performance degradation as data volumes grow. The PROC NETWORK step which performs network graph theory algorithms in SAS leverages standard node and link style table structures stored in the SAS CAS in-memory server. Techniques in this paper use a hybrid approach to manage graph data due to the CAS data input requirements of PROC NETWORK and the ability of the SAS data step to perform any necessary data transformations. Since PROC NETWORK uses node and link style datasets which must reside in CAS, data can be sourced and transformed through traditional table structures such as a relational database or SAS datasets, eliminating the additional storage layer considerations of a graph database. Any transformations or data manipulation can occur using SAS programming techniques and perform extremely fast due to SAS CAS in-memory server capabilities.

Data can be physically managed through many techniques. Processes described in this paper use SAS as a compute engine for traditional ETL processing from source data into a target data warehouse and for developing network analytics using the SAS in-memory CAS server. All data is physically stored in a traditional relational database using principles of graph theory to manage.

Figure 2 below provides a conceptual representation of how specific SAS technology can interact with different data sources.

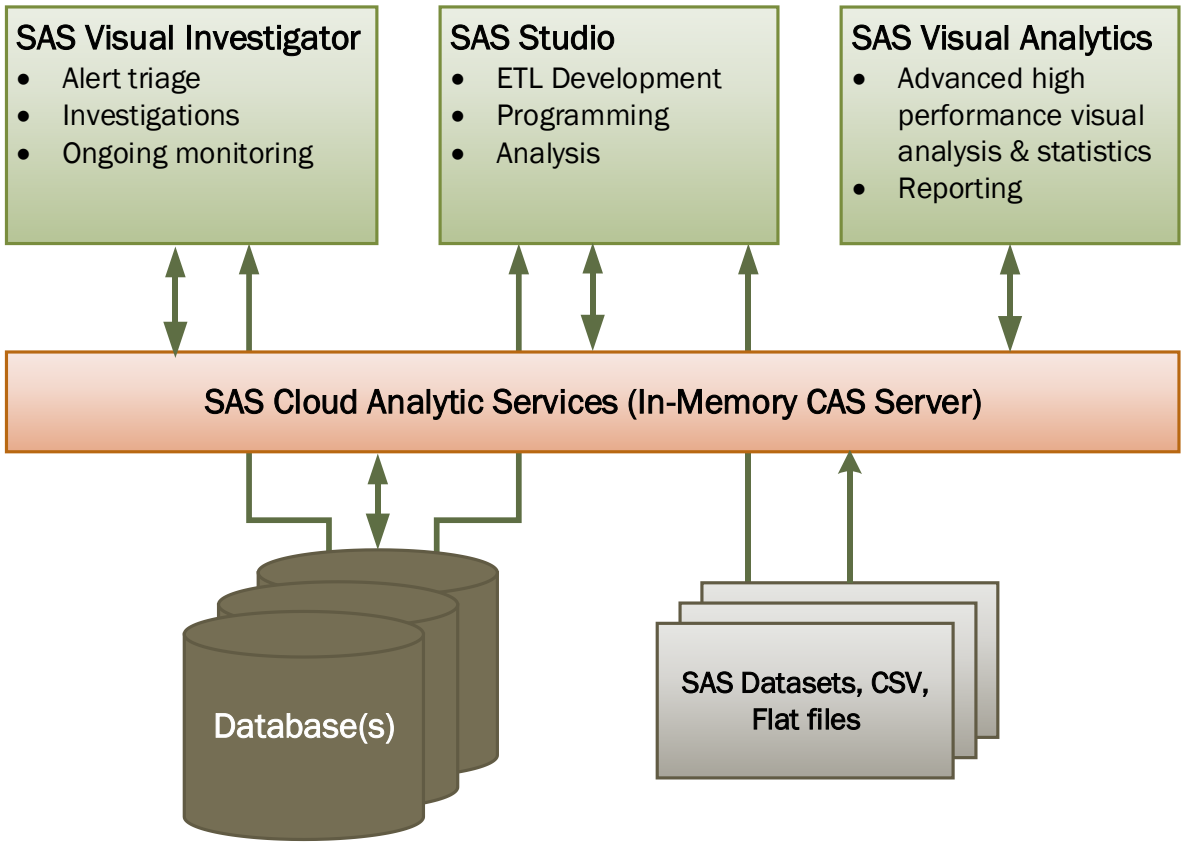

Figure 2: Multiple tools to manage data directly or indirectly through SAS Cloud Analytic Services.

See the References and Recommended Reading section of this paper for further background information regarding graph database technology and SAS Support documentation for different programming techniques and product documentation.

CONCEPTS TO MEASURE RISK THROUGH OPPORTUNITY

This paper quantifies risk through connections a customer may have through relationships to other entities, objects, events or any data which can describe possible means or opportunity to perform an activity being monitored. The data model described in this paper is used to demonstrate how corporate infrastructure can be used to measure the means and opportunity of individuals to move, hide or launder money. Data can be derived from both external sources of information and internal databases due to the flexibility of network linking. The data model consists of the following conceptual node types stored as dimension tables.

- Corporate registry information
  - Corporate Entities such as LLCs, S-Corps, Non-profits
  - Corporate Officers such as individuals or other corporate entities
  - Locations for officers or registered addresses for corporate entities
- Watchlist data representing risks necessary to monitor
  - Examples include marijuana-related entities, money service businesses, human trafficking risk or sanctioned entities
- Internal data such as customer or account information, transactions, events or other document filings

Corporate registry information and watchlist data is externally sourced while internal data is obtained from core systems of record. Each object can be modeled using type 2 slowly changing dimensional techniques for audit and version control if desired.

Further information regarding slowly changing dimension processing is available in the Recommended Reading section of this paper.

Dimension tables are used to store what are referred to as nodes in graph databases. Link tables are used to store what are referred to as edges in graph databases. Link records connect nodes across dimension tables or within the same dimension table. Link tables can also be referred to as bridge tables in a traditional relational data model.

A basic example of dimension tables in relation to a link table stored in a relational database is presented in Figure 3 below.

Corporate Entity (corp_entity)

GUID	Name	Type
1	Company A	LLC
2	Company B	S-CORP
3	Company C	Non-Profit

Corporate Links (link_corp)

From GUID	To GUID
4	1
5	2
6	2
7	3

Corporate Officer (corp_officer)

GUID	Name	Type
4	John Doe	Agent
5	Jane Doe	Manager
6	Josh Doe	Partner
7	Steve Doe	VP

Figure 3: Dimension tables with related link table cross referencing global unique IDs (GUIDs).

Global Unique IDs (GUIDs) are universal across all node dimensions in the data warehouse. GUIDs are stored as serialized integers which increment within and across dimensions. GUIDs are intended to be a systematic way to uniquely identify nodes across the network at an atomic level and not present a business-friendly identifier. Attributes stored in dimension tables and link tables are analogous to attributes stored on node and links within graph databases. While the physical structure of a link table may seem simple, the logical relationships which exist within data may present a more complex view.

Figure 4 below shows an example of how corporate structures can be stored with four physical tables and three soft relationships on the left; but have a much different structure represented by the data in the graph model on the right.

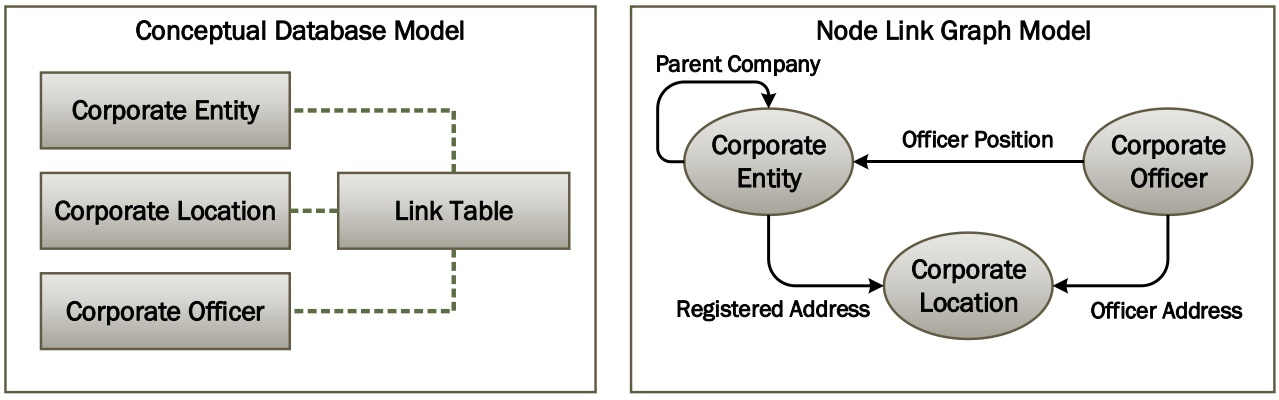

Figure 4: Traditional conceptual database model compared to graph data model represented by data.

Using a traditional relational database to physically store node and link style tables may be rigid but does help enforce standardization and controls within the data model itself. GUIDs are indexed appropriately for performance. As new objects are added, additional dimension and link tables can be created to capture necessary information about the object or entity as well as relationships which exist within the new subject area. Dimension and link tables

are intended to be very modular and primarily connected within individual subject areas. The following section describes techniques to connect disparate subject areas using processes such as entity resolution or geocoding address locations.

Enriching Data Using Analytics

Network analytics and other routines can be performed after central business data is stored across respective dimension and link tables. The following list provides example processes which are critical to highlighting risk as described in this paper.

- **Discovery models** establish artificial connections formed through entity resolution, geocoding or other processes. These are represented as explicit nodes with directed links to connections discovered in order to simplify the network graph model.
- **Centrality statistics** describe the structure of networks and entities within networks using statistics such as eigenvector, betweenness, closeness and degree counts.
- **Clustering** establishes physical boundaries of how nodes are connected. For example, if one node has the same cluster ID as another node, they are connected in some manner.
- **Community detection** can provide hierarchical logical groupings within clusters that tighten the lens around closely connected objects and improve BY variable processing.
- **Watchlist distances** help quantify risk through possible direct or indirect relationships to other bad actors or nodes of interest.
- **Reach counts** provide additional levels of risk by counting the number of risk elements within the reach of a given node.
- **Cycle detection** helps discover potential circular paths which can present risk through hidden connections.
- **Eccentricity** measures the longest shortest path and is useful in measuring the number of relationship layers, especially with corporate structures which have nested shell companies.
- **Risk scoring models** provide numeric attributes which can be used to filter, alert, analyze, screen or visually cue investigators and analysts.

Each process described in the list above can use SAS programming in SAS Studio to produce results using inputs from the core dimensional data. Output is stored within the same data warehouse and consumed by SAS tools such as SAS Visual Investigator and SAS Visual Analytics.

LAYERS OF INTELLIGENCE

Figure 5 below presents an overall conceptual layout of corporate registry tables as described above plus other core business dimensional objects such as watchlist data and internal data. Green objects represent discovery models which establish new artificial links between disparate sources of data. Orange objects represent output from multiple analytical processes. Additional analytical processes can be developed and stored as additional node dimensions as requirements arise.

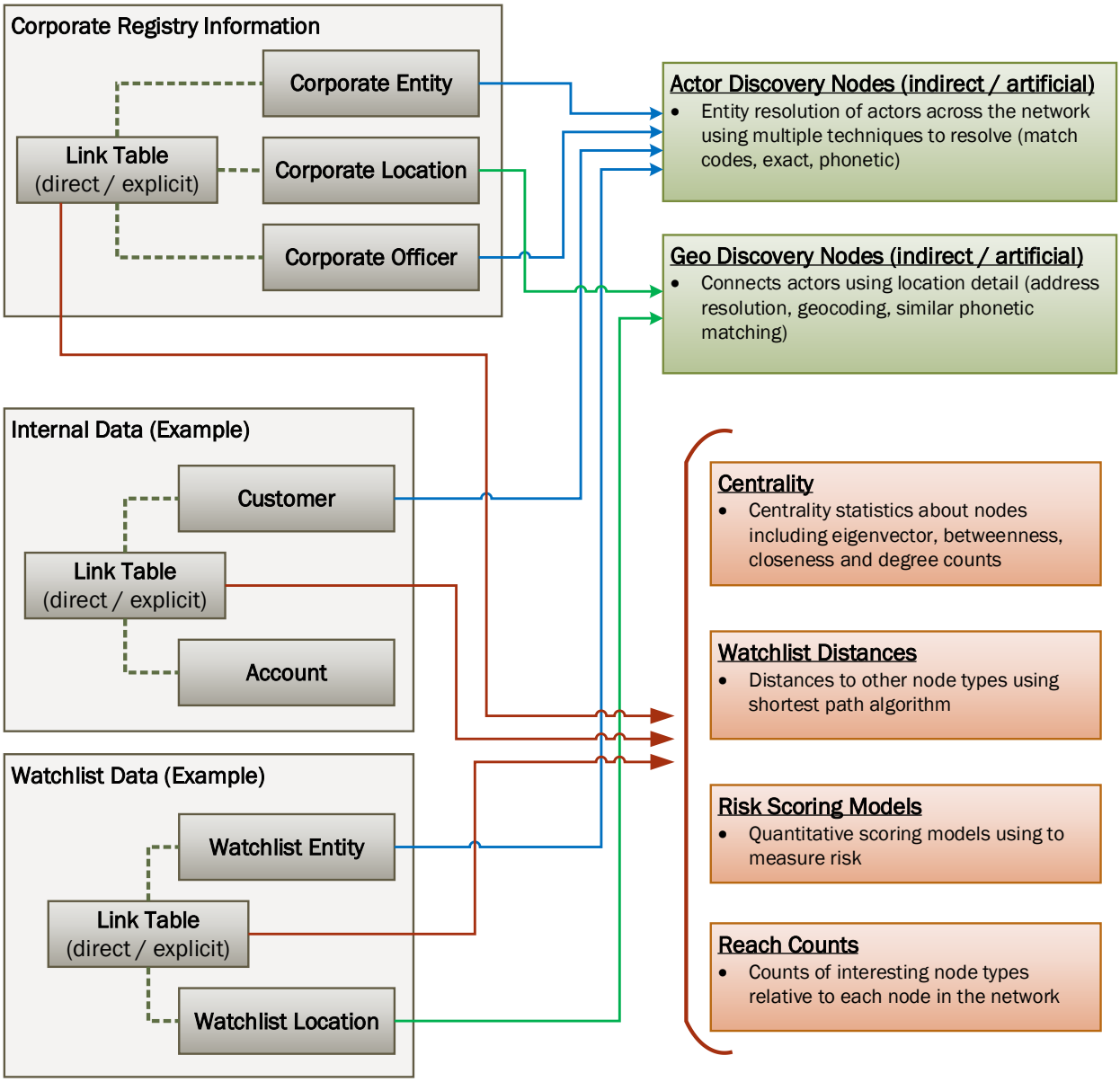

Figure 5: Core dimensional objects plus conceptual objects which store network analytics output.

ENHANCE RISK MODELS USING NETWORK ANALYTICS

Risk can be modeled many ways using SAS network analytics. The number of dimensions, attributes, network structures and overall techniques used in network analytics will vary from environment to environment. This paper will focus on a few examples related to quantifying the opportunity or ability to commit financial crime through legal corporate structures.

IDENTIFYING RISK USING CORPORATE ANALYTICS

Corporate analytics can be one route to take when measuring and quantifying risk based on a person or entities ability to commit a financial crime. Corporate entities are public sources of information in most countries. Corporate entities can be nested and layered around the world to create a complex web of hidden assets or channel to move, hide or launder money. Just because a person has a complex web of corporate entities does not make them a criminal; however, it does present more risk to the bank or financial institution due to the

number of potential indirect relationships that can exist as more layers of relationships unfold.

SAS Global Forum Paper 1786-2018 presents approaches to compute centrality, cycle detection, eccentricity, community detection, and clustering using the Panama Papers as example data containing corporate structures. SAS Global Forum Paper 1786-2018 also describes an approach for building a scoring model to incorporate network analytics. This paper will provide two additional risk dimensions using watchlist distance and reach counts with PROC NETWORK. Any of the risk metrics described can be stored and used for alert generation or to filter results in search and discovery in SAS Visual Investigator.

Compute Shortest Path Watchlist Distance

It may be useful to understand how close a non-watchlist node is to a watchlist node. Watchlist nodes can be important risks such as marijuana related entities, money service businesses, entities within the Panama Papers or other items of interest that may present risk. The shortest path algorithm within PROC NETWORK can provide an integer number which reflects the number of links between a node and the closest target node of interest. In the example SAS source code provided below, the following steps are taken.

1. Load links across the network into CAS.
2. Extract nodes from the links table.
3. Prune links to focus on confident connections only.
4. Isolate and identify target watchlist items.
5. Define super sink node concept and explicit link weight to improve processing times.
6. Execute PROC NETWORK to compute distances from all source nodes to target artificial super sink node.
7. Finalize output data and incorporate into permanent data output.

Sample SAS source code to compute watchlist distances using the steps above is provided below:

```
%load_links_cas(
    target_cas_lib=sixdcas,
    target_table=links,
    source_link_tables=
          core.link_corp
          core.link_icij
          core.link_mrb
          core.link_trafficking
          core.link_sanctions
          core.link_msb
          sixdegdm.link_geo_discovery
          sixdegdm.link_actor_discovery,
    node_types=Y,
    link_confidence=Y
);

/** Get nodes from link table **/
data sixdcas.nodes;
    set sixdcas.links(
           keep=from_guid from_node_type
           rename=(from_guid=guid from_node_type=node_type))
        sixdcas.links(
           keep=to_guid to_node_type
           rename=(to_guid=guid to_node_type=node_type))
```

```
    ;
    by guid;
    if first.guid;
run;

/** Only keep links where link_confidence >= 90 **/
data sixdcas.links;
    set sixdcas.links(in=links);
    if links and link_confidence >= 90 then output;
    drop link_confidence;
run;

/** Isolate watchlist nodes **/
data sixdcas.mrb_short_path_nodes;
    set sixdcas.nodes;
    if node_type in ('MRB_ENTITY', 'MRB_LOCATION') then
          source=1;
    else
          sink=1;
run;

/** Create super sink node links and define link weight **/
data sixdcas.mrb_links_super_src(keep=from_guid to_guid);
    set sixdcas.nodes(keep=guid node_type rename=(guid=from_guid));
    if node_type in ('MRB_ENTITY', 'MRB_LOCATION') then
          do;
                to_guid=-1;
                output;
          end;
run;
data sixdcas.mrb_links;
    set sixdcas.links sixdcas.mrb_links_super_src;
    weight = 1;
run;

/** For each source S (all nodes), find all the sinks T (super source 1
degree from watchlist nodes) for which there is a path from S to T and
return distance as weights **/

proc network
    links = sixdcas.mrb_links
    nodes = sixdcas.nodes
    nodesSubset = sixdcas.mrb_short_path_nodes;
    linksVar
          from = from_guid
          to = to_guid
          weight = weight;
    nodesVar
          node = guid;
    nodesSubsetVar
          node = guid;
    shortestPath
          source = -1
          outWeights = sixdcas.mrb_path_weights;
run;
```

```
/** Redistribute data across grid since super source method puts everything
on a single worker **/
data sixdcas.mrb_path_weights(partition=(sink) orderby=(sink));
    set sixdcas.mrb_path_weights(drop=source);
run;

/** Finalize output data **/
data sixdcas.mrb_distance;
    set sixdcas.mrb_path_weights(
          rename=(sink=guid path_weight=mrb_distance));
    by guid;
    mrb_distance = mrb_distance - 1;
    /* Subtract 1 because of super source */
    if mrb_distance > 0 then output;
run;
```

Data transformation is a critical piece of leveraging PROC NETWORK in CAS. If node or link data is not structured properly, DATA step programming can be used to transform data as needed.

Compute Reach Count to Watchlist Nodes

Reach is another useful function of PROC NETWORK which provides all nodes that are within a specified number of links or hops of a given node. Determining the number of watchlist items, jurisdictions or any item which may present risk within a specified max distance provides a more comprehensive statistic relative to watchlist distance described in the section above. Varying counts of watchlist nodes can be used to generate alerts within SAS Visual Investigator or to filter results within search and discovery.

In the example SAS source code provided below, the following steps are taken.

1. Load links across the network into CAS.
2. Extract nodes from the links table.
3. Prune links to focus on confident connections only.
4. Filter corporate entities which have links to analyze for improved performance.
5. Define surrogate reach identifier to compute reach distances for every single node.
6. Execute PROC NETWORK to determine all GUIDs within 4 degrees of every record in dim_corp_entity.

Sample SAS source code to compute reach counts using the steps above is provided below:

```
%load_links_cas(
    target_cas_lib=sixdcas,
    target_table=links,
    source_link_tables=
          core.link_corp
          core.link_icij
          core.link_mrb
          core.link_trafficking
          core.link_msb
          sixdegdm.link_geo_discovery
          sixdegdm.link_actor_discovery,
    link_confidence=Y,
    node_types=Y,
    relationship_type=Y
);
```

```
/** Get nodes from link table **/
data sixdcas.nodes;
    set sixdcas.links(
          keep=from_guid from_node_type
          rename=(from_guid=guid from_node_type=node_type))
        sixdcas.links(
          keep=to_guid to_node_type
          rename=(to_guid=guid to_node_type=node_type))
    ;
    by guid;
    if first.guid;
run;

/** Only keep links where link_confidence >= 90 **/
data sixdcas.links;
    set sixdcas.links(in=links);
    if links and link_confidence >= 90 then output;
    drop link_confidence;
run;

/** NOTE: Only keep corporate entities which have links **/
data sixdcas.dim_corp_entity;
    merge
          sixdcas.dim_corp_entity(in=base)
          sixdcas.links(in=to_links keep=to_guid rename=(to_guid=node))
    ;
    by node;
    if first.node and base and to_links then output;
run;

/** NOTE: Need reach identifier to be serial integer **/
data sixdcas.dim_corp_entity / single=yes;
    set  sixdcas.dim_corp_entity;
    length reach 8;
    reach = _N_;
run;

/** Compute Max Reach 4 degrees **/
proc network
    links     = sixdcas.links
    nodesSubset = sixdcas.dim_corp_entity;
    linksvar
          from = from_guid
          to = to_guid;
    reach
          maxReach = 4
          outReachNodes = sixdcas.corp_entity_reach_nodes;
run;
```

PROC NETWORK in the sample code above outputs all GUIDs which are within 4 degrees of every corporate entity. There are many methods to count the number of node types. DATA step programming can be used to count node types and define explicit cumulative totals which can be stored permanently and used for additional risk scoring or for display purposes in SAS Visual Investigator.

DISCOVER RISK PATHWAYS USING SAS VISUAL INVESTIGATOR

Node decorators in SAS Visual Investigator can be used to highlight potential risk pathways from non-watchlist nodes or nodes which are considered to be internal customer data. Node watchlist distances can be used to define values within objects which are used to display icons on network diagram nodes.

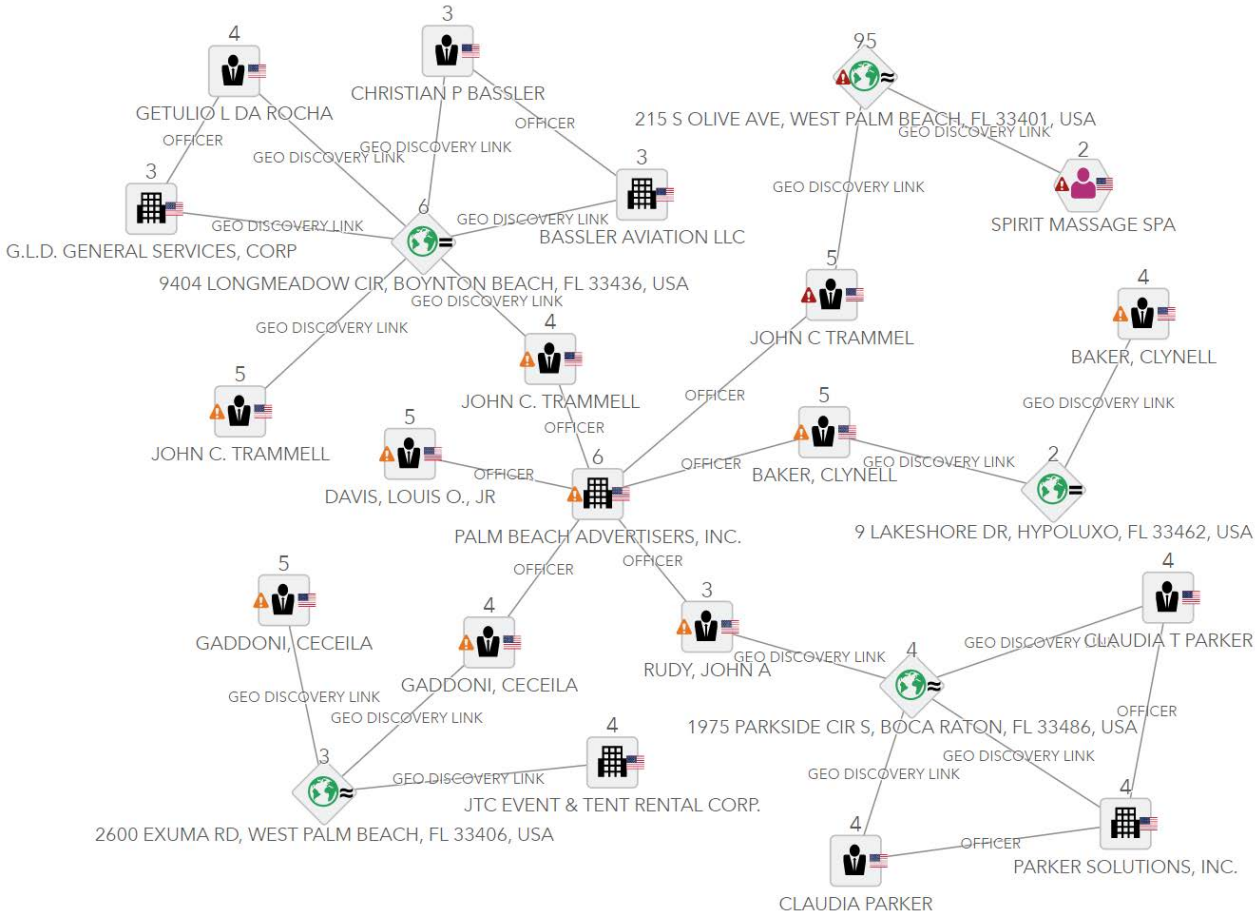

Figure 6: Example network diagram in SAS Visual Investigator with risk pathways highlighted using node decorators.

Figure 6 above presents an example network of corporate entities with a corporate officer in the upper right corner near a human trafficking risk node. Nodes within 4 degrees of the human trafficking risk node in the upper right corner have an orange exclamation icon displayed on the left side of the node. Nodes within 2 degrees of the human trafficking risk node have a red exclamation icon. The purpose of this technique is to promote productivity for the investigator by guiding the network growth down paths which will eventually present risk.

SAS Visual Investigator version 10.6 provides additional ways to dynamically format and color nodes and links themselves for even more flexibility with providing visual cues for investigators.

CONFIGURE NODE DECORATORS IN SAS VISUAL INVESTIGATOR

SAS Visual Investigator object administration uses JSON to configure node decorators. Example JSON used to display node decorations in Figure 6 is provided below:

```
{
    "svi_link_count": {"text": "{{svi_link_count}}", "position": "N"},
    "wl_dist_risk_band": {"src": "{{wl_dist_risk_band}}", "position": "W"},
     "jurisdiction_country_code": {"src": "{{jurisdiction_country_code}}",
     "position": "E"}
}
```

The example node decorator JSON code above assumes the following columns exist within the source table.

- **svi_link_count** displays the link count above nodes for performance reasons. This is computed ahead of time using the degree metric provided by PROC NETWORK.
- **wl_dist_risk_band** displays an image with the filename of either HIGH, MED, or LOW based on values in the source table. This is calculated from watchlist distance values provided by PROC NETWORK described in the respective section above.
- **jurisdiction_country_code** displays an image of the country flag where the jurisdiction is located based on values within the source table.

A screenshot of network node annotation configuration using JSON in SAS Visual Investigator is shown below in Figure 7. Clicking the question mark icon in the upper right corner of the Network node annotation textbox within SAS Visual Investigator provides additional usage notes as well.

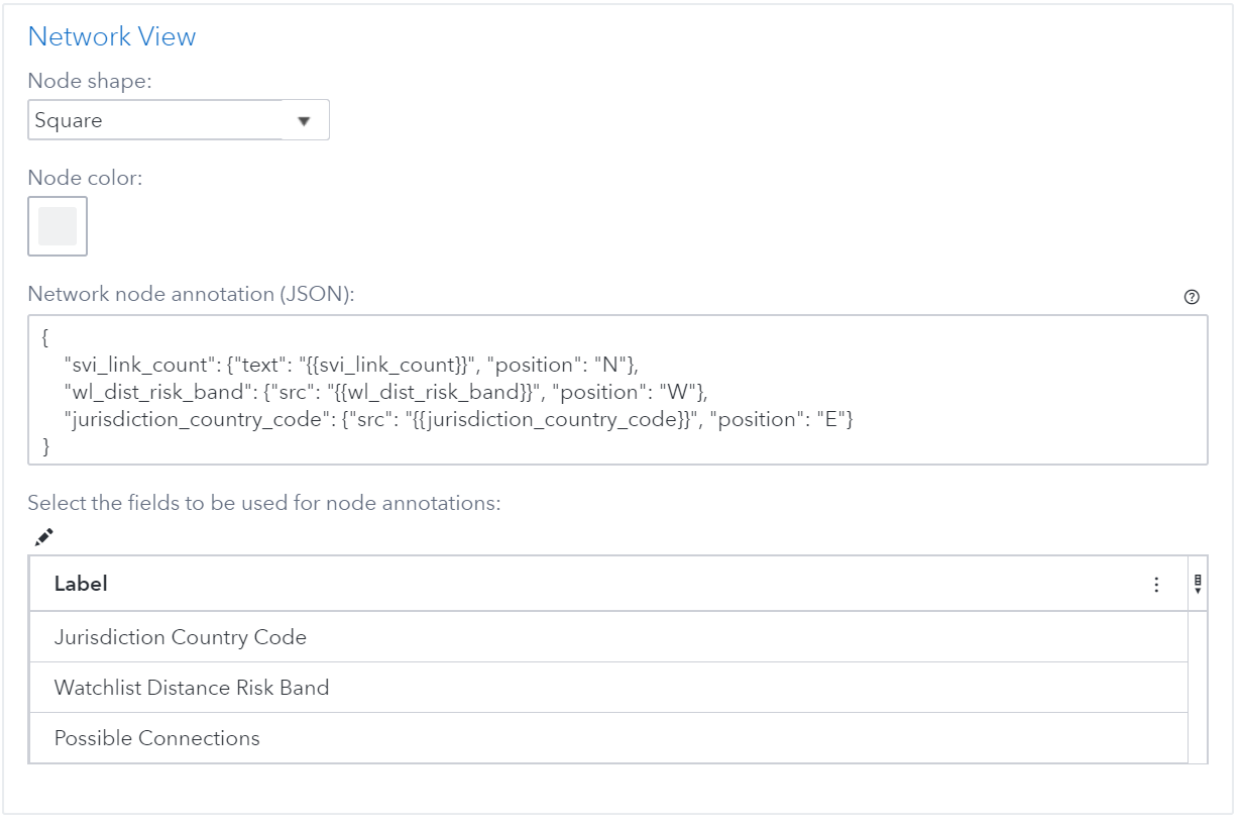

Figure 7: SAS Visual Investigator node decorator configuration.

CONCLUSION

Risk can be managed many ways and present many challenges for organizations of all sizes. SAS can be used to measure, analyze, improve and monitor if the appropriate tools are made available to data scientists. SAS technology leveraged within this paper includes many products within SAS Viya such as SAS Studio, SAS Visual Data Mining and Machine Learning and SAS Visual Investigator. Criminals will always find ways to penetrate the financial system for their own gain. For this reason, managing risk is a journey and never a final destination.

REFERENCES

"Data Modeling Concepts and Techniques | Neo4J." Neo4J, Inc. Accessed February 26, 2020. Available at https://neo4j.com/developer/guide-data-modeling/.

Ryan, Dan. "FinCEN: Know Your Customer Requirements." Harvard Law School Forum on Corporate Governance. Accessed February 26, 2020. Available at https://corpgov.law.harvard.edu/2016/02/07/fincen-know-your-customer-requirements/.

"SAS Visual Data Mining and Machine Learning 8.2: The NETWORK Procedure." SAS Institute. Accessed February 26, 2020. Available at https://documentation.sas.com/?docsetId=casmlnetwork&docsetTarget=procnetwork_network_toc.htm&docsetVersion=8.1&locale=en.

"FFIEC BSA / AML Examination Manual." Federal Financial Institutions Examination Council. Accessed February 26, 2020. Available at https://bsaaml.ffiec.gov/manual.

Overton, Stephen. April 9, 2018. "Discovering Insightful Relationships inside the Panama Papers Using SAS Visual Analytics." Proceedings of the SAS Global Forum 2018 Conference. Available at https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1786-2018.pdf.

"Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management." Office of the Comptroller of the Currency. Accessed February 26, 2020. Available at https://www.occ.gov/news-issuances/bulletins/2011/bulletin-2011-12.html

RECOMMENDED READING

- *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*
- *Networks: An Introduction*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Steve Overton
6 Degree Intelligence
(919) 341-9667
stephen.overton@gmail.com
https://www.linkedin.com/in/overton/