Paper 1786-2018

Discovering Insightful Relationships inside the Panama Papers Using SAS® Visual Analytics

Stephen Overton, Overton Technologies

ABSTRACT

Network analytics is a broad methodology which supports the desire to perform link analysis through visual tools such as SAS® Visual Analytics, SAS® Social Network Analysis, and SAS® Visual Investigator. Link analysis visually displays all possible relationships which exist between entities, based on data available, to provide insight into direct and indirect associations. This can be a very helpful tool to support an investigation as a part of a fraud or anti-money laundering investigation process. Beneath the surface, data management techniques and advanced analytical routines are used to discover relationships, transform data and build the appropriate data structures to support link analysis. Network statistics can help describe networks more accurately by using quantitative data to define complexities and unusual connections between entities. This paper will explore an approach to support network analytics and link analysis by using the Panama Papers as a real-world example. The Panama Papers leak is the largest leak of confidential data to-date. The data contained within the Panama Papers provides a wealth of knowledge to financial investigation units because it exposes previously unknown relationships between corporate entities and individuals.

INTRODUCTION

Analyzing networks of social relationships can be challenging because of the complex connections which can branch multiple directions and grow many layers deep. This can occur as little as a few degrees of separation, up to dozens of layers which span thousands of links. The average human mind can only perceive a handful of these associations at a given time without some assistance such as a network graph visualization. In some cases, the hidden link which establishes a complex network could be the critical point in a financial crime investigation. This paper will demonstrate methods to aid in the consumption and assimilation of complex network data through data management and visualization techniques using SAS® Visual Analytics running on the SAS® Viya platform with SAS® Visual Data Mining and Machine Learning utilized for NETWORK procedure programming capabilities.

This paper is focused towards the analysis of potentially suspicious behavior revolving around activity such as money laundering, tax evasion, fraud, bribery, corporate beneficiary ownership restructuring or other criminal activity facilitated through layering techniques to obscure the true source of activity or beneficiary. The targeted reader of this paper is any type of analyst involved in the assessment, exploration, monitoring, or investigation of network data similar to the data presented in this paper.

One major advantage gained from network analysis is increased productivity and effectiveness of a risk-based methodology for prioritizing, ranking, and scoring an organization's customer base. Network analysis data and decisions made from it could potentially be an additional risk dimension in an organization's model risk management practices. Generically speaking, this risk dimension could be representative of a customer's opportunity to commit financial crime or obfuscate criminal activity through different layers of relationships.

DISCLAIMER

Data used in this paper was obtained from the International Consortium of Investigative Journalists (ICIJ) under the Open Database License (ODbL) version 1.0 and the Creative Commons Attribution-ShareAlike license. For more information on the source of the Panama Papers and related Offshore Leaks data, as well as additional licensing details, see the References section below.

Examples given in this paper and presentation do not suggest or imply criminal intent. There can be legitimate use for offshore companies and trusts. Not every person or corporation in the ICIJ Offshore Leaks has broken the law or acted improperly. The examples given in this paper and supporting presentation are chosen because of the visual appeal for demonstration purposes.

EVERYTHING STARTS AND ENDS WITH THE DATA

Sound data management practices should be the cornerstone of every analytical solution. This section will briefly cover key concepts to define the data management strategy used to organize information and share beneficial metrics used in the analysis of the Panama Papers and other related data in the ICIJ Offshore Leaks. These metrics will be used to visualize information more effectively in SAS® Visual Analytics. These metrics can also be used similarly in other tools which visualize relationships such as SAS® Visual Investigator.

CONCEPTUAL DATA MODEL

Network data at the most basic level is comprised of two elements, nodes and links. Nodes can represent a range of actors or objects such as corporate entities, loan applications, phone numbers, vehicles or people. Links represent connections between nodes. Links are represented at the minimum by two variables which provide what node identifier the link is "from" and "to". Both nodes and links at the atomic level can have many attributes for classification, filtering, and analysis purposes. Additional tables can be defined to organize nodes and links more effectively and provide aggregate level information.

The following diagram shown in Figure 1 provides a conceptual view of nodes, links, and other tables used to facilitate network analytics.

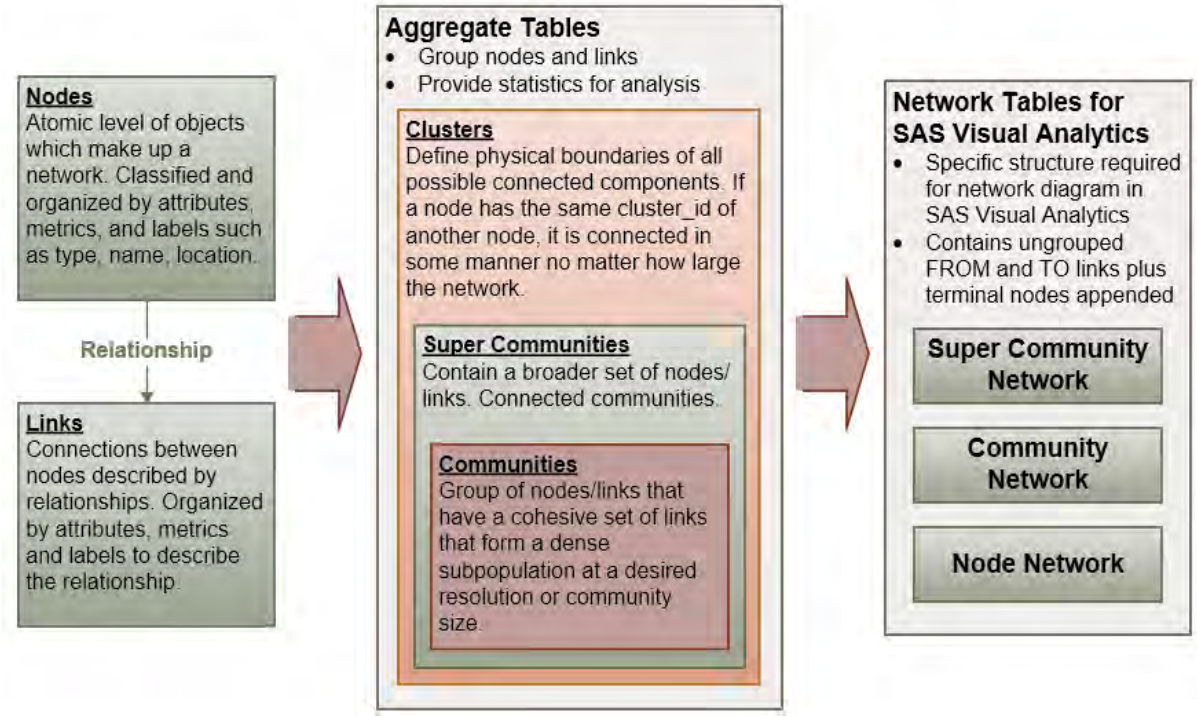


Figure 1: Conceptual Network Data Structures Used to Facilitate Network Analysis.

The following diagram shown in Figure 2 provides a summary of the relationships defined between the different node types of the ICIJ Offshore Leaks as February 2018. Node types within the ICIJ Offshore Leaks can be a physical address, corporate entity, intermediary of the corporate entity, officer of the corporate entity, or a custom holding group. Holding groups are manually identified relationships provided by the ICIJ.

Relationship Type	Frequency
CONNECTED_TO	12,145
INTERMEDIARY_OF	594,692
MANUAL	9
OFFICER_OF	1,667,166
PROBABLY_SAME_OFFICER_AS	132
REGISTERED_ADDRESS	768,200
SAME_ADDRESS_AS	5
SAME_AS	166
SAME_COMPANY_AS	15,523
SAME_ID_AS	3,120
SAME_INTERMEDIARY_AS	4
SAME_NAME_AS	79,929
SIMILAR_COMPANY_AS	203
TERMINAL	981,260
UNDERLYING	1,238

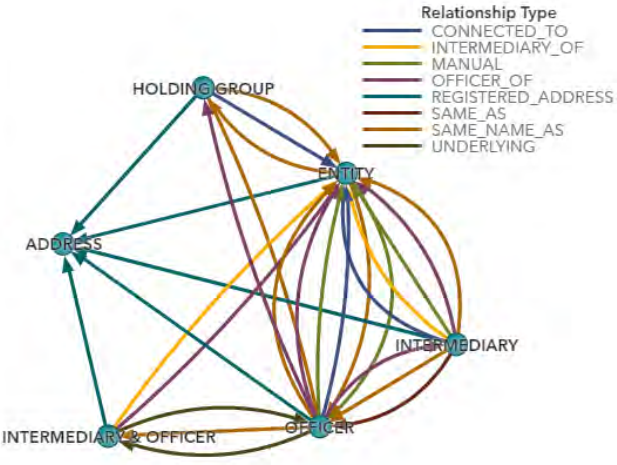


Figure 2: Summary of Node Types and Relationships.

Network Direction

Network direction is another critical factor to define when analyzing relationships and processing network data. The direction of a network refers to the links between nodes, and is either directed or undirected. Directed links can be used to describe a specific type of relationship between two nodes in a more declarative sense while undirected links can be used for general associations between nodes. Directed network types can have a bi-directional undirected relationship if there are opposing links between two nodes. Arrows are used to visually designate the direction of the link in directed network types.

In the context of financial crime investigations, one assumption that can be made is that if a person, entity, or officer has a link, the communication can flow both directions and the associated risk applies regardless of direction. If information and communication can flow both directions, an undirected network type is the best definition for network direction. Another broad assumption that could be made is that links between nodes are investigation leads at the minimum, requiring human confirmation of the relationship to minimize the chance of missing suspicious activity. If this assumption is true, treating the network as undirected provides the proper context for the investigation process and data management processing. In general, the context of data analyzed and business workflow should determine the definition and usage of network direction.

DATA MANAGEMENT PROCESS

The following section describes the high-level approach used to prepare the ICIJ Offshore Leaks data for network analysis in SAS® Visual Analytics. SAS programming was developed using SAS® Studio. Key pieces of SAS code are provided and briefly explained in the context of the analysis demonstrated in this paper. SAS Support documentation can be referenced for detailed syntax explanation and additional information on the NETWORK procedure. The References section at the bottom of this paper provides a URL to obtain the broader set of SAS programs used to manage the ICIJ Offshore Leaks data.

The following diagram shown in Figure 3 provides a high-level summary of how source data is processed for network analysis in SAS® Visual Analytics.

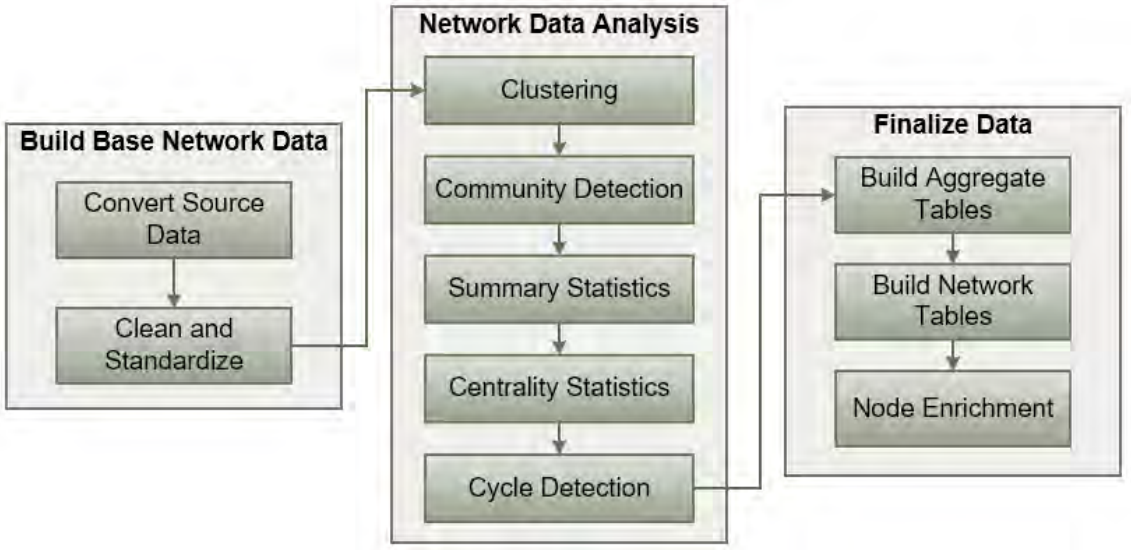


Figure 3: High-level summary of data management process.

Obtain Source Data

The ICIJ has been collecting data and leaks from many sources and media partners since 2013. A wealth of knowledge and intelligence has been put into organizing the data in an effective network data model for public consumption. Source data used in this paper can be obtained directly from the ICIJ website. The URL for this website can be found in the References section below. Data is provided for each Offshore Leak source. Further details on how the ICIJ obtained the actual source data can be referenced from the ICIJ website.

- Bahamas Leaks: Contains information from the corporate register of the Bahamas.
- Offshore Leaks: Initial data added in June 2013 as a part of the ICIJ 2013 Offshore Leaks expose. This data provides information from two offshore service providers.
- Panama Papers: Data leaked from the Panama law firm Mossack Fonseca. Largest leak of confidential data to-date. The Panama Papers provides the first public glimpse of how offshore corporate entities are structured around the world.
- Paradise Papers: Data leaked from another offshore law firm, exposing similar corporate structures described in the Panama Papers.

Establishing Links between Nodes

Links between nodes are provided by the investigative work performed by the ICIJ as well as their media partners. This is extremely helpful for analyzing data from the Panama Papers and other sources within the Offshore Leaks. Other data sources may not have links defined explicitly. Defining links between nodes is a critical task to complete in any form of network analysis.

Convert Source Data to SAS Data Sets

As of February 2018, node and link data from the ICIJ is provided as comma-separated values in plaintext files. Data was converted to data sets initially using manually written SAS programming, but later enhanced to use SAS® Enterprise Guide import functionality to better handle elements within the source data such as carriage returns within fields.

Example SAS code used to convert CSV files to SAS data sets:

```
data papers.watchlist;
    infile '/sasdata/offshore_leaks/data/source/watchlist.csv'
    delimiter = ',' MISSOVER DSD lrecl=32767 firstobs=2 encoding='UTF-8';
```

```
    informat node_id 20.;         /* numeric */
    informat node_name $200.;     /* character */
    format node_id 20.;
    format node_name $200.;
    input
          node_id ??  /* numeric */
          node_name $ /* character */
    ;
run;
```

Clean and Standardize Data

Source data is rarely delivered perfectly. One key advantage of SAS® Viya is the ability to transform data and take necessary steps to cleanse for data quality purposes using SAS programming. After the ICIJ source data is converted to SAS data sets, the data is standardized using basic techniques to capitalize certain fields, clean extraneous characters, and build a more generalized node data set. Country codes are also translated to full country names for display purposes. Developing this process helps gain a greater understanding of the source data context as well.

Two key findings discovered in the cleansing and standardization of the ICIJ Offshore Leaks data:

- Duplicate links can exist between nodes because of multiple relationship types that can exist between nodes. For example, a person can serve as the original creator of a corporate entity as well as the CEO or other registered agent.
- Some nodes can exist as an intermediary and an officer. Steps are taken to define a special node type to distinguish these nodes without losing context.

Cluster Nodes and Links to Establish Boundaries

The first critical element of organizing node and link data is to define the maximum distance any node can be related to another node, especially when working with large volumes of data. Clustering nodes and links into unique groups enables faster data processing because BY variable techniques can be leveraged to focus computations on individual segments of nodes and links, rather than having to scan the entire network for each computational step. Visual link analysis is more effective because nodes and links can be subset into the most logical representation using a single cluster reference. For example, if node A is never connected to node Z, either directly or through indirect linkages, then a bottom-up analysis does not need to consider node Z and a top-down analysis can itemize by each cluster for an initial segmented analysis.

SAS programming code used to define network clusters:

```
proc network
    links      = mycas.links
    direction  = undirected
    outnodes   = mycas.nodes_cluster_id
    outlinks   = mycas.links_cluster_id;
    linksvar
          from = from_node_id
          to   = to_node_id;
    connectedcomponents;    /* Connected Components algorithm */
run;
```

Check SAS Support documentation provided in the References section for more details on the Connected Components algorithm within the NETWORK procedure.

Define Communities and Super Communities within Clusters

Community detection provides greater depth to grouping nodes and links. When used in conjunction with clustering, a hierarchy of vantage points can be defined to group nodes and links at different resolutions. Community detection can also be performed multiple times on top of communities already detected to create additional layers of super communities or subcommunities which naturally cascade through layers of different resolutions. Communities and clusters provide a framework to promote more productive analysis and data processing because high volumes of nodes and links can be grouped into a logical drill path. SAS Support documentation can be referenced for more details on the specific mathematical routines used for community detection within the NETWORK procedure.

Super Communities

Community detection summarizes specific data into more general groups using very specific mathematical routines. Super communities are formed by using the link output from an initial community detection analysis to form communities on top of communities. In other words, a super community is a group of communities. This can be performed as many times as necessary to form additional summarizations of nodes and communities at different levels of resolution. The analysis performed as a part of this paper only defines a layer of communities which logically group nodes and links, as well as a layer of super communities which logically group communities.

Figure 4 provides an example report from SAS® Visual Analytics which shows a network diagram of super community relationships on the left and a filtered view of the communities within a selected super community on the right.

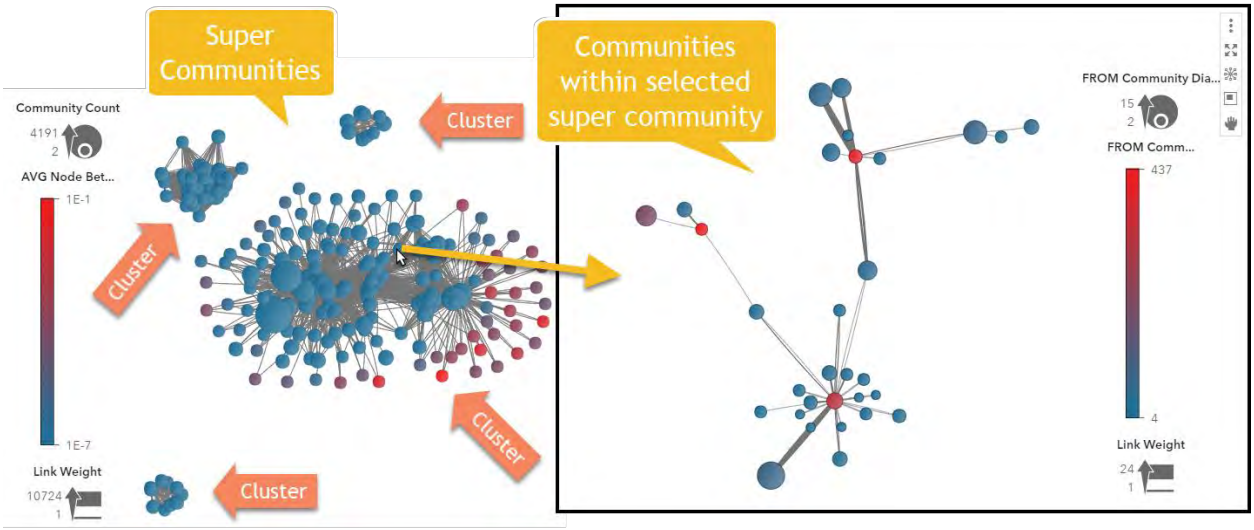


Figure 4: Community View linked from Super Community Network View.

Clusters define absolute boundaries between nodes and links. Communities are more focused than clusters and will have links between the communities defined if the criteria used to define communities permits.

SAS programming code used to define network communities:

```
proc network
    links        = mycas.links_aggregated
    direction    = undirected
    outnodes     = mycas.nodes_community_all
```

```
    outlinks     = mycas.links_community_all;
    linksvar
          from   = from_node_id
          to     = to_node_id
          weight = weight;
    community                    /* Community Detection Algorithm */
          recursive(maxcommsize = 2500)
          outlevel     = mycas.community_summary
          outoverlap   = mycas.community_overlap_all
          outcommunity = mycas.community_node_count_all;
run;
```

Critical elements of the community algorithm shown above:

- Output data defined by the “outlinks” parameter identifies communities at the most atomic link level, which can be summarized to determine what communities are connected to one another by merging the community ID for both the “from” and “to” nodes.
- The “recursive” option forces the community algorithm step to continually break down communities until they are below the “maxcommsize” parameter. Communities which are symmetric in nature are not broken down and notated in the log. This is helpful and productive for the analysis of the ICIJ Offshore Leaks because a symmetric network is most likely a star-shaped network, and therefore does not need to be broken apart. A maximum community size of 2500 is used because this approaches a size limitation of the network analysis data visualization in Visual Analytics. Realistically it also becomes very cumbersome to visually analyze 2500 nodes in a single visualization. Therefore, this number should be further reduced to optimize visual ergonomics with effective grouping of nodes and links.
- The “weight” option provides input for the community algorithm to consider some links strong than others. The input data set “links_aggregated” aggregates the multiple relationship types that can exist into a count of links between two nodes. While most nodes have a link weight of 1, some can have link weights greater than 1.
- Output data defined by the “outoverlap” parameter provides information on what nodes overlap between more than one community. The NETWORK procedure provides a primary community ID for the individual node, but does not assign a community ID for the links which reference the node from other communities. In other words, links which do not have a community ID are links which bridge two different communities. Understanding and storing the number of communities a node is a member of provides another useful metric for finding what generally holds a network together and forms critical connections between communities.

Check SAS Support documentation provided in the References section for more details on the community detection algorithm within the NETWORK procedure.

Compute Network Summary Statistics

Summary statistics describe networks of data in terms which can help quantify the complexities of how nodes and links are related. Summary statistics can be generated for nodes, links, and the population analyzed by the NETWORK procedure. If BY variable processing is being used, each BY group is summarized in the output table defined by the “out” parameter as shown in the example SAS programming code below.

SAS programming code used to compute network statistics:

```
proc network
    links     = mycas.links(where=(link_community_id>0))
    direction = undirected
    outnodes  = mycas.nodes_sum_stats_by_community
    outlinks  = mycas.links_sum_stats_by_community;
    linksvar
```

```
          from = from_node_id
          to   = to_node_id;
    by link_community_id;
    summary
          connectedComponents
          shortestpath = unweight
          out          = mycas.community_summary_stats;
run;
```

The following summary statistics were effective in analyzing the ICIJ Offshore Leaks, in order of importance:

- Community Diameter: Integer number representing the longest shortest path in a network of nodes and links. Helps quantify the width or depth of a community or other population of network data. Larger numbers are indicative of layers of complex relationships which increase the opportunity for information to flow across many nodes within each layer of the community.
- Node Eccentricity: Like diameter, eccentricity is an integer number representing the longest of all shortest paths from an individual node's perspective. Very useful in describing how many possible degrees of separation exist from an individual node. Larger numbers represent a higher degree of risk due to the depth of relationships an individual node contains.
- Isolated Stars: Binary flag representing if the network analyzed is a star shape. Many of the largest communities of nodes and links within the ICIJ Offshore Leaks are star-shaped, due to intermediaries which register many corporate entities.

Check SAS Support documentation provided in the References section for more details on all statistics computed as a part of the Summary statement within the NETWORK procedure.

Compute Centrality Statistics

Centrality statistics are more advanced than summary statistics because they describe the interconnectedness of nodes and links in relation to one another. Centrality statistics can be used to further describe the complexities within network data using quantitative metrics.

SAS programming code used to compute centrality statistics:

```
proc network
    links         = mycas.links_aggregated(where=(super_community_id > 0))
    direction     = undirected
    outnodes      = mycas.nodes_centrality_by_community
    outlinks      = mycas.links_centrality_by_community;
    linksvar
          from    = from_node_id
          to      = to_node_id
          weight  = weight;
    centrality
          clustering_coef
          close   = weight
          between = weight
          eigen   = weight
          degree;
    by super_community_id;  /* Performing at super community level */
run;
```

The following centrality statistics were effective in analyzing the ICIJ offshore leaks, in order of importance:

- Betweenness: Metric which describes the number of times a node or link occurs on a shortest path

within a network. Very useful in quantitatively describing gatekeepers of information and communication flow because of the relative position in a network. A higher betweenness score indicates a more probable path of communication.

- Eigenvector: Uses a centrality importance scoring mechanism to compute the importance of nodes in a network based on their neighboring nodes. This metric is useful to describe actors such as key influencers, important facilitators of suspicious activity, or popularity in a network.
- Closeness: Metric which describes how fast information can spread from a node to other nodes in the network based on distance. A higher node score means that node is closer to other nodes, relative to the other nodes in that network.
- Degree: Integer number representing the number of connections to a node. Useful measure for displaying node size in a network diagram.

Check SAS Support documentation provided in the References section for more details on Centrality statistics within the NETWORK procedure.

Example Risk Score Methodology Using Network Analytics

A node and link risk score can be calculated using a combination of centrality statistics and other descriptive statistics. This score can be leveraged by visual network diagrams as roles such as node or link size and color.

The following SAS programming DATA step code snippet provides an example of how centrality statistics can be combined with community overlap count to form a node risk score:

```
/* Standardize risk score to be used on network visualizations */
    node_risk_score =
        sum(1, (centr_between_wt*100), (centr_eigen_wt*100),
        (centr_close_wt*100), (5 * sqrt(coalesce(community_count,1))));
```

The node risk score above leverages betweenness, eigenvector, closeness, and community overlap count to create an overall risk score because these were found to be productive indicators of risk. In this example, centrality statistics are weighted the same and multiplied by 100 to form a possible range of 1 to 100. Community overlap count is normalized using the square root function and weighted 5 times higher than centrality statistics.

Detect Cycles within Communities

Cycle detection is a powerful tool that helps identify when a node is indirectly related to itself through at least one other node. Cycle length is the number of links which define a cycle within a network. A very low cycle length is common among mutual relationships such as shared account holders. As the cycle length grows, so does the complexity and risk. In financial crime, cycles can be used to detect things such as fraud rings, nexuses, or opportunities to layer illicit activities. Determining what cycle length is suspicious requires feedback from actual investigations and analysis to define the optimal point of what cycle length provides the most productive investigation.

SAS programming code used for cycle detection:

```
proc network
    direction       = undirected
    links           = mycas.links_aggregated(where=(link_community_id>0));
    linksvar
          from      = from_node_id
          to        = to_node_id;
    by link_community_id;
    cycle
          out       = mycas.cycles
          minlength = 5;
run;
```

```
proc fedsql sessref=casauto;
    create table community_cycle_length{options replace=true} as
          select
                link_community_id as community_id,
                max("order") as cycle_length
          from cycles
          group by link_community_id
    ;
quit;
```

Check SAS Support documentation provided in the References section for more details on the Cycle detection algorithm within the NETWORK procedure.

The following SAS® Visual Analytics report shown in Figure 5 provides a visual summary of cycles which exists within a complex network contained in the ICIJ Offshore Leaks.

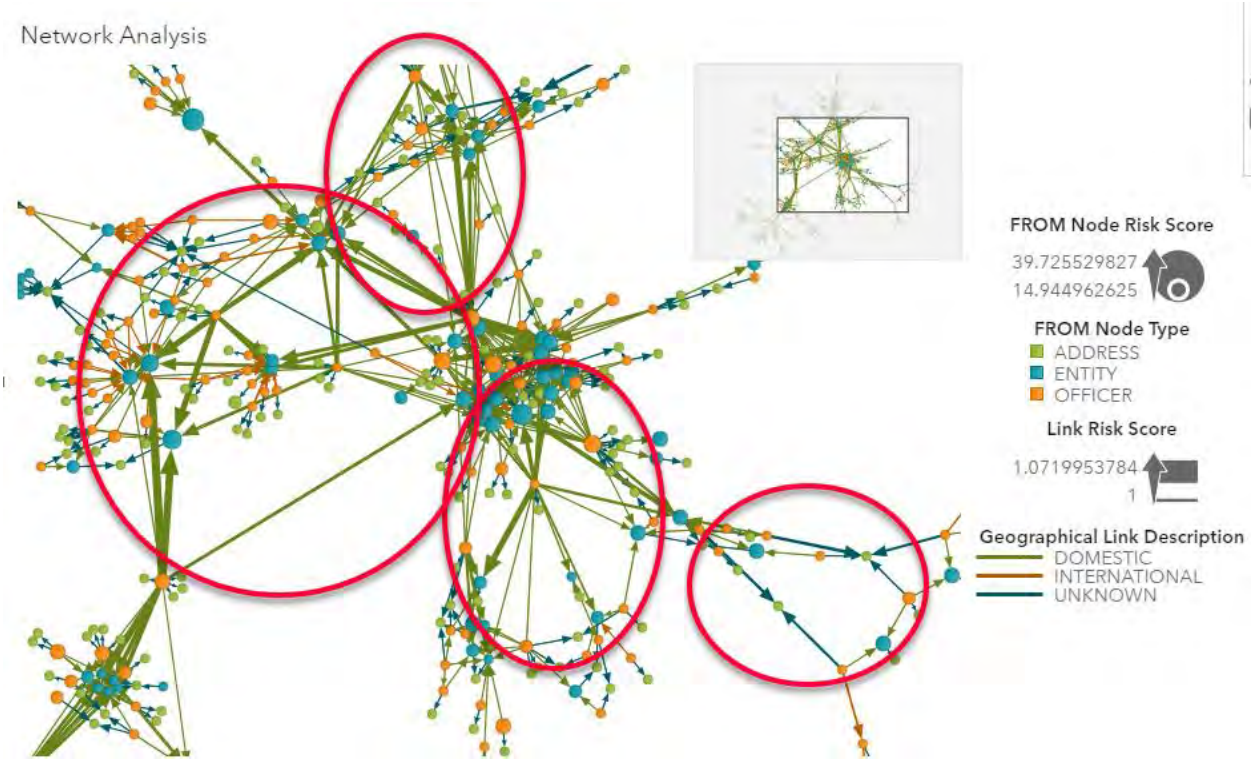


Figure 5: Example Network Diagram showing cycle detection.

Build Aggregate Tables

After network analysis is performed, primarily using the NETWORK procedure, data is finalized for SAS® Visual Analytics exploration. Aggregate tables which summarize clusters, communities and super communities are produced using SQL and DATA step programming to join descriptive summary statistics into a single view per network object. Other useful metrics are added to provide productive metrics that highlight counts of high risk nodes or activities contained within the respective community or cluster. Example high risk metrics based on the ICIJ Offshore Leaks include the count of Bearer shares, country count, intermediary and entity count. These metrics act as additional variables to consider when filtering down communities or clusters for top-down analysis.

Build SAS® Visual Analytics Network Tables

As of SAS® Visual Analytics version 8.2, network diagrams support two different input table structures to perform visual analysis. The ungrouped, or paired-node structure is used in this paper because of the

robust atomic level of detail. To build network tables for SAS® Visual Analytics, terminal nodes must be extracted and appended to the original table of links. In addition, node attributes and statistics are joined for both the "from" node and the "to" node. Network tables are built for the original nodes and links provided by the ICIJ, as well as communities and super communities. Any final formatting can be accomplished in this step as well.

For further information on SAS® Visual Analytics and the usage of network diagrams, refer to the SAS® Visual Analytics User Guide provided in the References section below.

Enrich Nodes with Parent Object Statistics

The final step copies essential aggregate statistics and joins to the atomic node table to provide context for bottom-up analysis based on top-level data. Summary statistics at the node-level can also promote predictive modeling analysis by providing key response variables to analyze the correlation of node variables.

PERFORMING LINK ANALYSIS OF THE PANAMA PAPERS

Analyzing network data can be a cumbersome task due to the complexities and many directions that can unfold as relationships are followed and researched. Each link and node can lead to extended amounts of research to validate the purpose and understand behavior. In financial crime monitoring and investigations, link analysis can be used in two key ways:

1. Supporting an investigation of an alerted entity or person by providing all possible leads through direct and indirect related entities. This can include any historical alerts, cases, or other watchlist feeds that may be related to the alerted entity or some other indirectly related entity within the network.
2. Monitoring and alerting on individuals or organizations based on a network of relationships rather than a single entity or directly related entity. Suspicious networks of individuals can be collected into an entire network to investigate for activities such as fraud rings. New customers or relationships can be screened against suspicious networks to monitor for related activity on incoming business.

The next two sections will provide methodologies to support link analysis from a bottom-up and top-down perspective using SAS® Visual Analytics and the data management process defined in previous sections of this paper.

TOP-DOWN ANALYSIS

Analyzing networks using a top-down approach starts by viewing or processing the entire population while leveraging clusters and communities to segment analysis and drill into areas of interest. Centrality and descriptive summary statistics can be used to guide the analysis down a desired path using multiple techniques. Visually performing this analysis in tools such as SAS® Visual Analytics leads to specific scenarios that can be developed using SAS programming to monitor in fraud detection scenarios. This section will walk through an example network analysis performed using the ICIJ Offshore Leaks.

Overall Summary of Nodes within the ICIJ Offshore Leaks

Analyzing the Panama Papers, Paradise Papers, and other related leaks can be a very interesting subject due to the big power players that exist within the data. Politicians, business owners, and celebrity names provide very topical examples that are easy to connect with and research. The following crosstab shown in Figure 6 provides a summary count of the different node types based on the network tables built within the data management process. Terminal nodes represent leaf nodes which have no outbound links to other nodes. It is interesting to note that physical addresses are mostly terminal points in the network.

Total Nodes by Type from Network Links

link_desc	LINK	TERMINAL	Total
FROM Node Type	Node Count	Node Count	Node Count
ADDRESS	5	374,934	374,939
ENTITY	306,588	478,352	784,940
HOLDING GROUP	2,919	.	2,919
INTERMEDIARY	24,570	.	24,570
INTERMEDIARY & OFFICER	1,139	.	1,139
OFFICER	719,573	28	719,601
Total	1,054,794	853,314	1,908,108

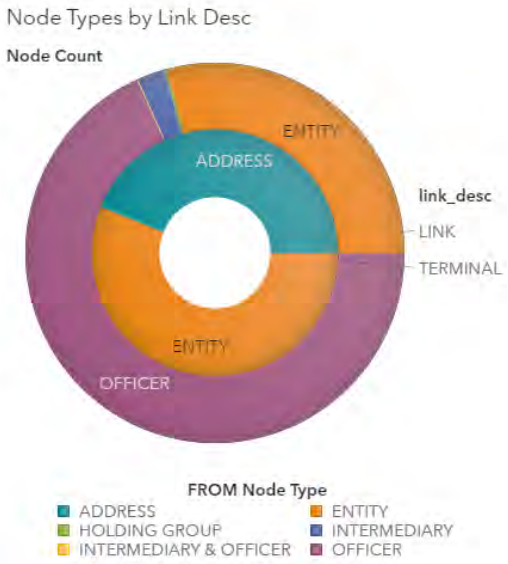


Figure 6: Summary count of node types.

The following crosstab shown in Figure 7 provides an overall summary of the sources of information from the ICIJ.

Total Links by Source

link_desc	LINK	TERMINAL	Total
Source ID	Node Count	Node Count	Node Count
(missing)	170,012	853,314	1,023,326
Bahamas Leaks	41,519	.	41,519
Offshore Leaks	201,519	.	201,519
Panama Papers	257,667	.	257,667
Paradise Papers - Appleby	103,374	.	103,374
Paradise Papers - Aruba corporate registry	80,932	.	80,932
Paradise Papers - Bahamas corporate registry	2,148	.	2,148
Paradise Papers - Barbados corporate registry	42	.	42
Paradise Papers - Cook Islands corporate registry	682	.	682
Paradise Papers - Lebanon corporate registry	1	.	1
Paradise Papers - Malta corporate registry	190,987	.	190,987
Paradise Papers - Samoa corporate registry	5,919	.	5,919
Total	1,054,794	853,314	1,908,108

Figure 7: Summary count of source ID provided by the ICIJ.

Cluster Summary

Cluster summarization is performed after understanding the overall population of node types as well as other secondary ad hoc analysis to get comfortable with the data. The following table in Figure 8 provides an overall distribution of the number of nodes contained in the largest clusters within the ICIJ Offshore Leaks data as of February 2018. It is interesting to note that prior to the release of the Paradise Papers, over 90% of nodes were all connected in a single cluster. As of the February 2018 release of the Paradise Papers, the distribution has started to spread out more as shown below in Figure 8 due to more information added outside of Cluster ID 1.

Cluster Summary

Cluster ID	Node Count	% Total Node Count	Hyper Community Count	Diameter	Community Count
1	933,109	48.89%	183	.	46,597
98	284,325	14.90%	47	.	7,831
49	161,754	8.48%	9	.	3,043
96	161,499	8.46%	10	.	1,565
106	72,375	3.79%	1	.	540
217	35,360	1.85%	1	.	1
180	16,632	0.87%	1	.	57
152	7,566	0.40%	1	.	70
214	4,640	0.24%	1	9	17

Figure 8: Cluster summary node distribution.

These top clusters shown in Figure 8 contain too many nodes to visually analyze. Super communities provide the next layer of analysis.

Super Community Visual Network Diagram

Using Cluster ID 1 as an example, the following super community view shown in Figure 9 provides a visual layout of all super communities contained within Cluster ID 1. Each super community is represented by a single node, which contains many communities within each node. Larger nodes represent a higher betweenness core, which represents communities with nodes that are more along critical paths within their respective super community. Darker shades of red indicate a higher volume of communities within the super community. It is interesting to analyze the shape and organization of communities within differing types of super communities, both quantitatively and visually.

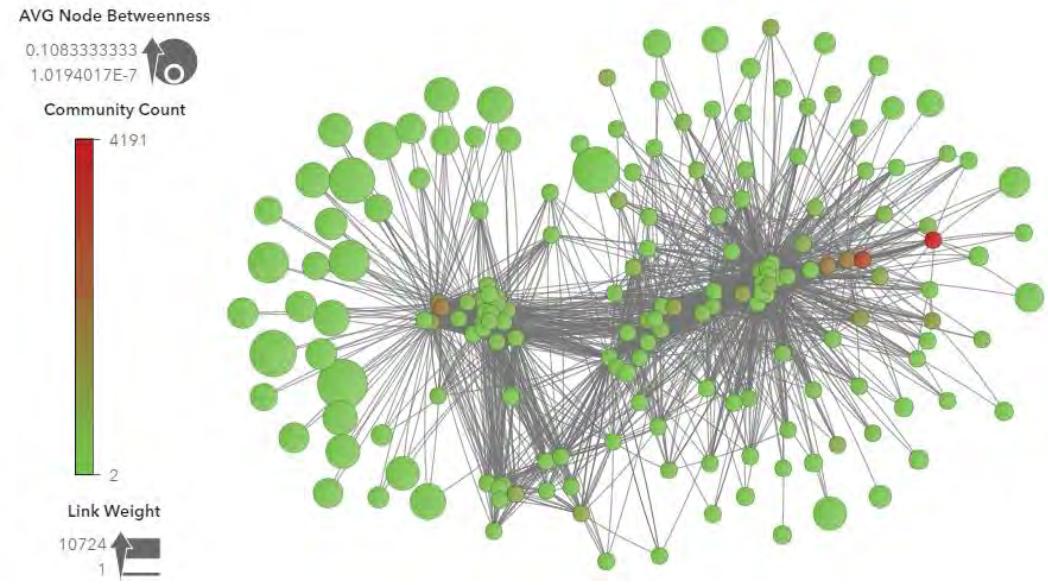


Figure 9: Cluster ID 1 super community network view.

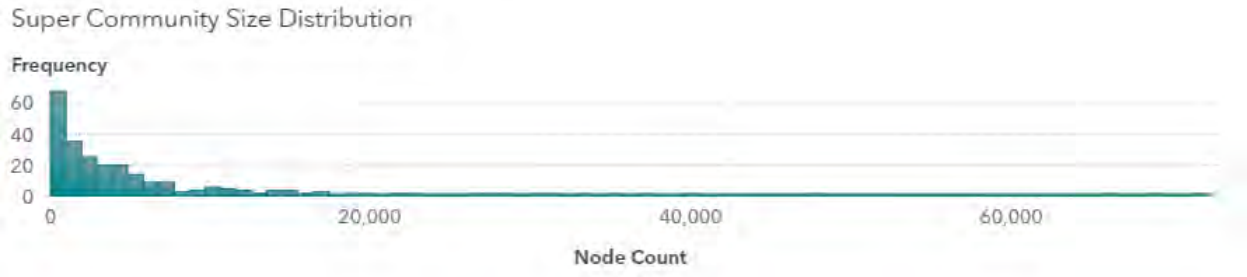
Visually the outer nodes shown in Figure 9 above have a higher average betweenness score, yet have a lower degree count. This means nodes are more likely to be on critical paths within the communities of these super communities, but there are fewer paths that flow between the super communities. Super community nodes within the middle of the network shown in Figure 9 have a higher degree count and a

lower average betweenness score, which leads to a possible assumption that the interconnectedness of super communities may be diluting the betweenness scores of nodes within those communities.

Super Community Statistical Summary

The following histogram and table shown below in Figure 10 provides a more detailed breakdown of the super community statistics within the network diagram shown in Figure 9 above. It is interesting to note the range of node counts across all super communities in this cluster. A higher community count of the super community does not necessarily mean the node count will be high. It is more likely that the specific count of intermediaries within the super community is more indicative of a higher node count because intermediaries establish the other node types. But node count does not necessarily indicate risk; node count is generally used to describe overall size and potential complexity. Another interesting observation to note is the extremely low ratio of intermediaries to other node types for super communities with the highest community count.

Other important risk metrics shown in Figure 10 below include the country count and bearer count. High country counts mean the network is spread across more jurisdictions which potentially creates more risk because of the global reach. Bearer shares represent physical pieces of paper that an individual possesses to officially own a corporate entity. Bearer shares are suspicious because they hide the actual owner of a business or firm.



Super Community Summary

Super Community ID	Cluster ID	Community Count	Node Count	AVG Community Diameter	Entity Count	Intermediary Count	Country Count	Bearer Count	AVG Node Risk Score
132	1	4,191	16,743	2.3655452159	4,830	8	77	9	13.865978352
187	1	3,252	15,707	2.6752767528	3,928	4	64	4	14.317522084
240	1	2,036	9,295	2.5913555992	2,698	9	45	1	16.060201898
63	1	1,906	10,172	2.870933893	2,965	4	81	189	14.493397444
135	1	1,808	12,122	2.6028761062	3,730	107	83	601	14.556691681
239	1	1,482	12,409	3.5020242915	4,601	1,062	74	159	12.206657813
238	1	1,481	11,552	3.4199864956	3,793	1,029	76	151	17.699923074
118	1	1,106	7,585	1.9104882459	4,340	12	47	2,638	15.520394793
140	1	1,046	7,221	2.5717017208	2,026	4	36	1,706	16.141505997

Figure 10: Super community node distribution histogram and summary statistics.

Deeper Dive into Community Level Analysis

As an example, Super Community ID 102 shown in Figure 11 is selected for deeper analysis of the community links within because of the higher ratio and volume of intermediaries to other node types, a very high Bearer Count, a higher than normal average community diameter, and a very high country count.

Super Community Summary

Super Community ID	Cluster ID	Community Count	Node Count	AVG Community Diameter	Entity Count	Intermediary Count	Country Count	Bearer Count	AVG Node Risk Score
123	1	786	4,259	2.1921119593	1,731	4	24	11	20.48540672
43	1	765	5,870	1.9529411765	4,219	1	46	4	21.188060487
8	1	729	4,112	1.9794238683	1,759	6	68	960	16.529034605
102	1	726	21,928	4.1831955923	11,730	873	116	3,641	12.631789422
103	1	712	4,079	1.1474719101	3,692	1	5	348	34.407722417
141	98	681	30,228	5.7797356828	8,349	.	140	.	13.061228762
97	1	652	4,353	1.9509202454	2,263	1	13	.	19.592857505

Figure 11: Example super community selection with summary statistics.

The following network diagram in Figure 12 provides a deeper breakdown of the communities within Super Community ID 102 shown above in Figure 11.

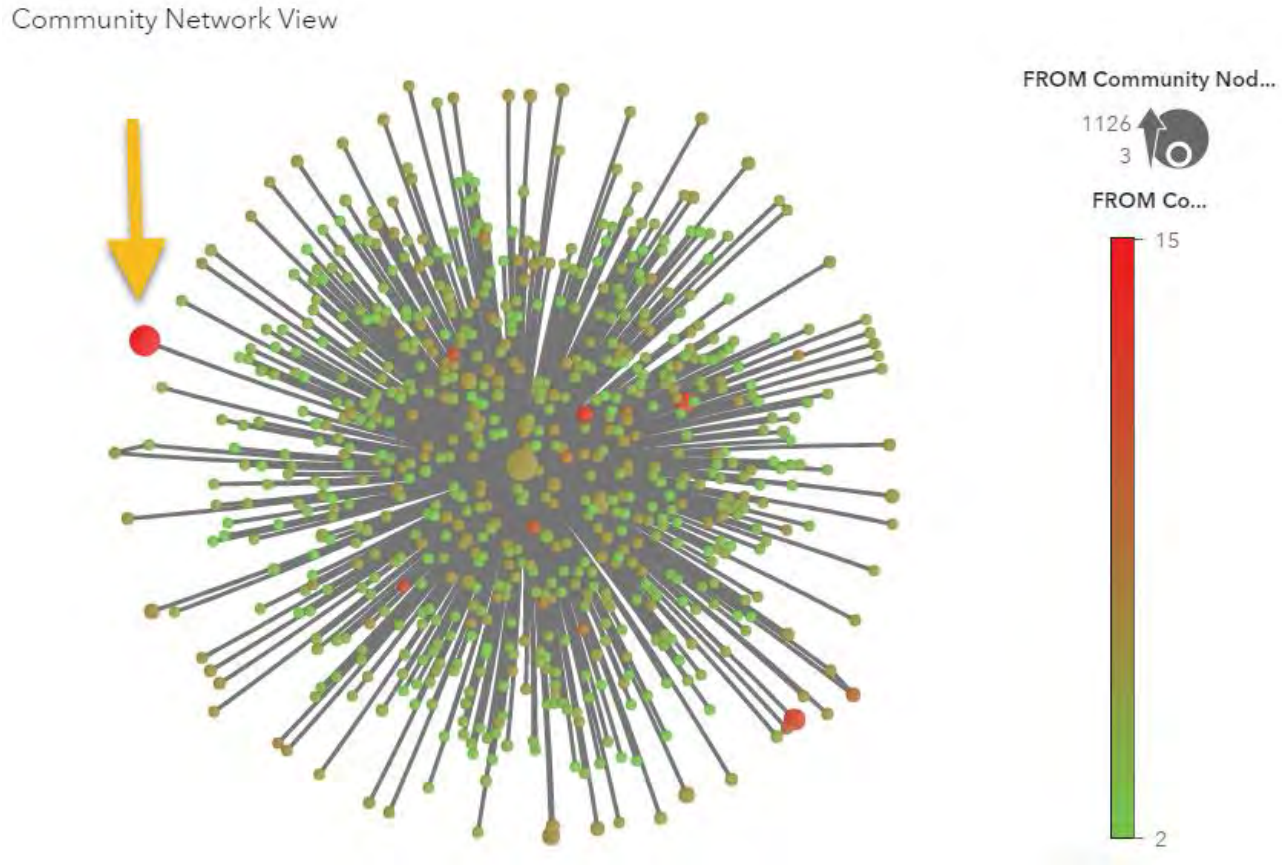


Figure 12: Community links within selected example super community.

The network diagram shown in Figure 12 provides an interesting view of communities within Super Community ID 102. Node size provides context to the volume of nodes within the community while node color represents the diameter of the community. It is interesting to note the outer node representing a high volume of nodes as well as a high community diameter. This community is selected for yet another drill down analysis as shown in Figure 13 below.

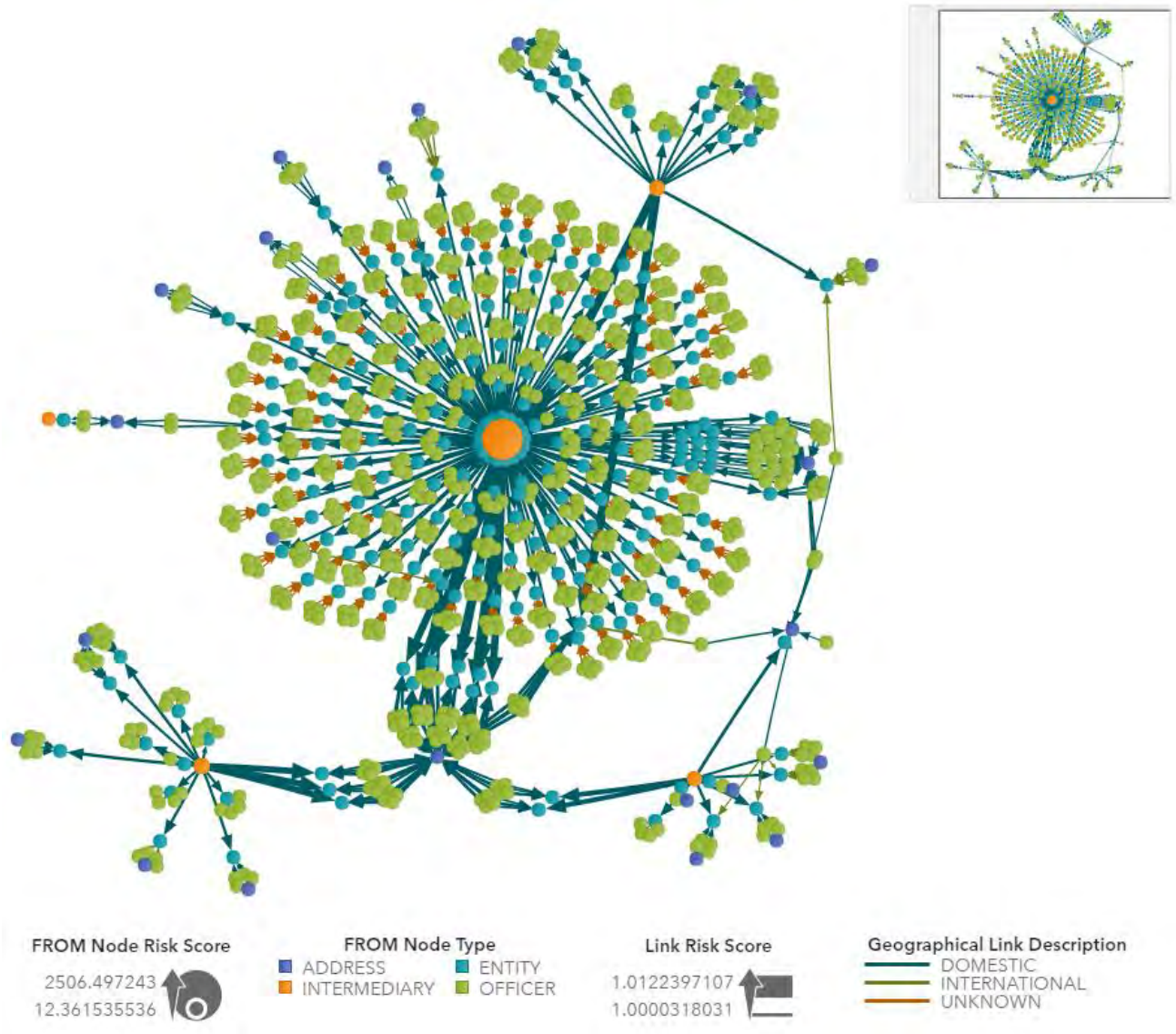


Figure 13: Network diagram for community of interest.

Conclusions from Top-Down Analysis

Figure 13 above provides the lowest level of the top-down approach for link analysis because individual nodes are shown. The actual names provided by the ICIJ Offshore Leaks are hidden for confidentiality purposes in this paper but are available in the source data. Key elements of this network diagram are the following:

1. Node size represents a Node Risk Score described in the previous Data Management Process section outlining Centrality Statistics. Larger nodes have a higher risk due to the connections they have within this community network along with an overlapping community count with other communities not shown.
2. Node color represents the Node Type as either a physical address, corporate entity, intermediary, or officer.
3. Link width represents the Link Risk Score described in the previous Data Management Process section outlining Centrality Statistics. Larger links represent a higher betweenness score.
4. Link color represents if the source and target node have different jurisdictions. International

relationships are riskier. Visual analysis confirmed the links are mostly domestic, but further analysis not shown here confirms that individual nodes are physically located across 5 different countries.

5. Visual analysis confirms a few cyclical relationships between the larger group of corporate entities and the smaller outlier groups.
6. Nodes which exist along cyclical relationships, 2nd or 3rd degree relationships from intermediaries that connect to other 2nd or 3rd degree relationships would be interesting to research as a part of an actual investigation.

From this point, the investigator could step through each node of the network diagram to research individual officers, corporate entities, and addresses to understand relationships and potentially gain insight into hidden purposes previously unknown. This type of information could lead to discovery of new motives by individuals under investigation as well as opportunities to commit financial crime or other illicit activities.

BOTTOM-UP ANALYSIS

Bottom-up analysis is a classic methodology used to investigate a suspicious person's network of related entities. Traditional bottom-up network analysis methods have been around for a long time, often dramatized by popular police detective TV shows using whiteboards or corkboards with push buttons and string along with mugshots of the persons of interest that surround a suspect. This process could be potentially replicated much faster with technology, partially at the minimum, at least to the degree of which data can be sourced to build out the network.

The pathway for a bottom-up analysis begins from an individual node and grows outward to community and potentially super community level analysis. The key driver that stops a visual network analysis is the volume of nodes and links displayed. Too many nodes can overwhelm the end user and become counterproductive. A specific node to begin from depends on the purpose and context of the analysis. An actual investigation could create the need to perform bottom-up analysis or an ad hoc exploration of a broader network analysis could spark the need to focus on a single node and build from the bottom-up. SAS® Visual Analytics does not have network growth functionality. Instead, layers of communities and other node and link attributes must be used to effectively filter down the population of network data.

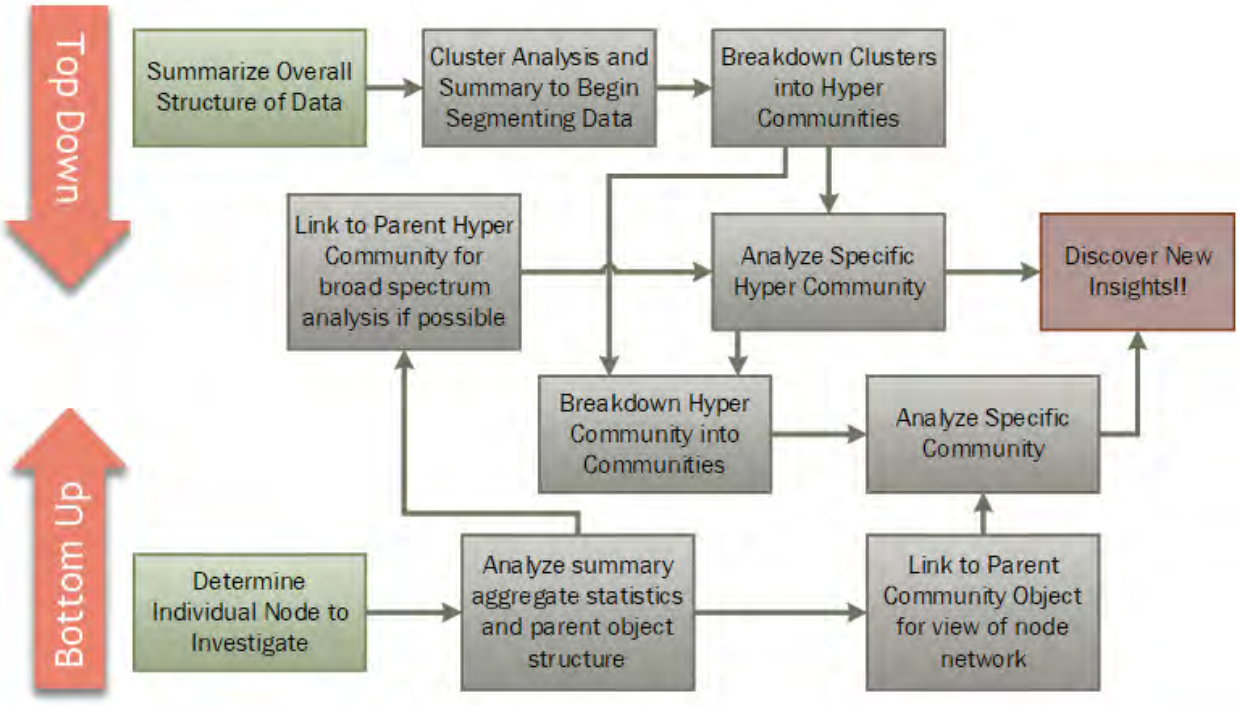


Figure 14: Top-down and bottom-up analysis workflow process.

The diagram shown above in Figure 14 summarizes the general workflow of a top-down and bottom-up analysis. The intersection of these two different analysis methodologies begins at the community level. Therefore, properly defining community resolution is critical for both analysis methodologies described in this paper. The top-down and bottom-up analysis can be considered "information pathways" as described in SAS Global Forum Paper 2960-2015. This paper can be referenced for more guidance and concepts around building an "information pathway" within SAS® Visual Analytics.

Establish Initial Node to Build Analysis Around

Filtering a single node is as simple as using a text entry box in Visual Analytics to filter down a list table of all nodes. A Parameter is used to store the text entered as a string, so the list table can leverage for fuzzy string matching. Node enrichment data is valuable for node-level data at this point because it gives top-level context from the bottom-up. Once the desired node is identified, interactions can be defined using SAS® Visual Analytics "Actions" to link the list table object to any number of other objects to take the next step up in the analysis. The respective Community ID or Super Community ID is leveraged in the linkage between objects.

Analyze Summary Statistics for Community and Super Community

Example summary statistics about the parent community and super community can be shown as simulated in Figure 15 below. The respective unique ID is leveraged for the link filtering.

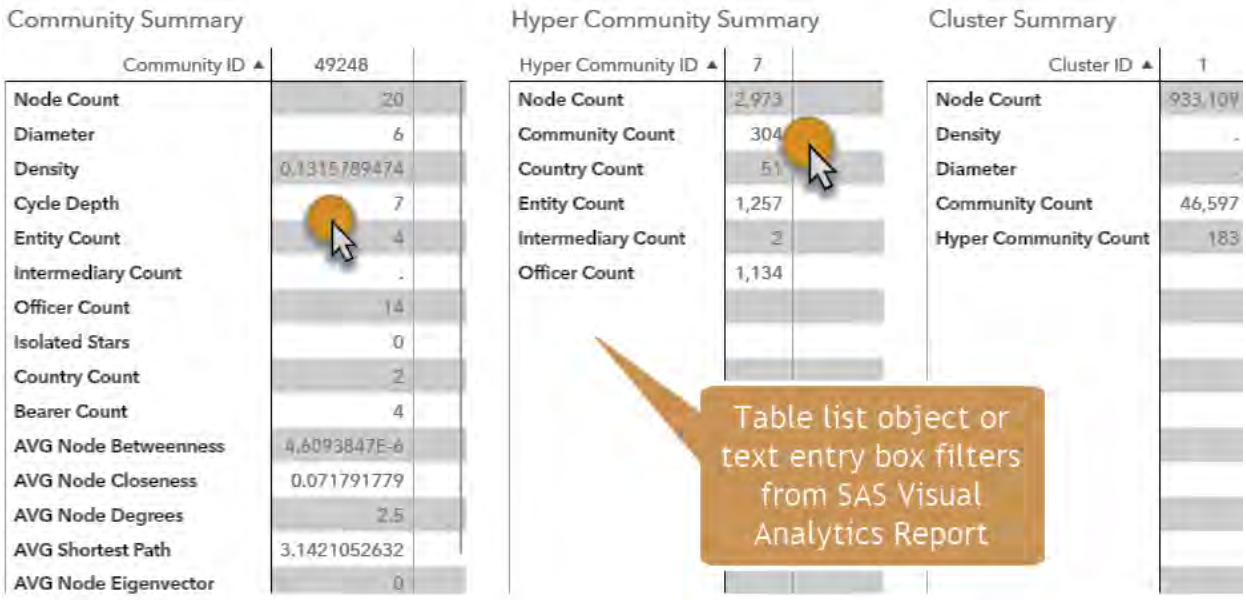


Figure 15: Example summary statistics for community and super community.

After understanding the complexity of the parent community, report or page linking can be used to pass the community ID into another view of the community using a network diagram. Additional objects such as the geomap can be used to visually understand spatial relationships.

Analyze Community or Super Community Network

A basic view of the community network is shown in Figure 16 below. This network is very similar to the same Power Player example provided on the ICIJ website for the base node used in this example, except a single node is missing because the NETWORK procedure decided it belonged in an adjacent community. Node labels are not shown in this example but can be enabled for an actual analysis. The network diagram can become overwhelming if there are too many node labels to display.

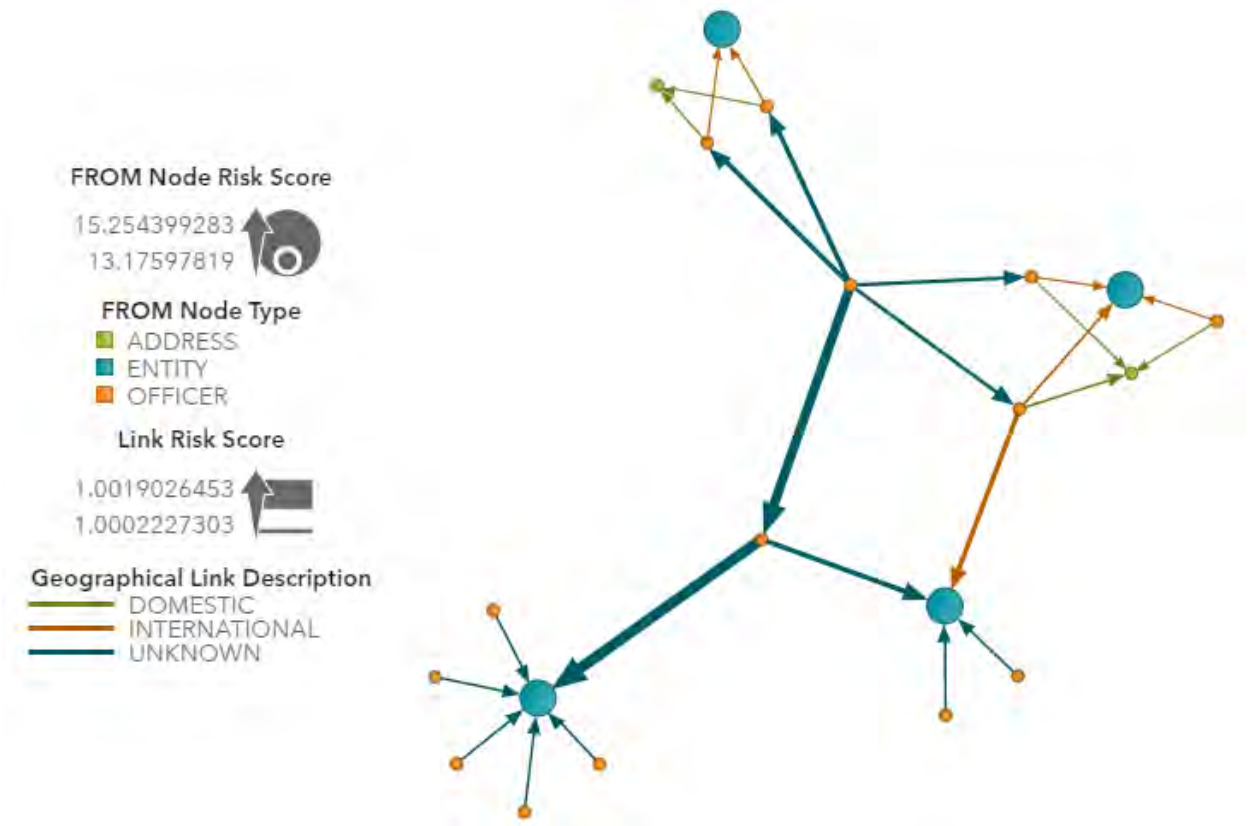


Figure 16: Community view of an ICIJ Power Player network.

The missing node from Figure 16 is a key intermediary in this example. The importance of this intermediary can be seen in the super community view of the broader network shown in Figure 17 below.

Conclusions from Bottom-Up Analysis

The intermediary removed from the community which contained the base node of the bottom-up analysis is a key facilitator in the much broader super community. The key intermediary is located in the middle of the network shown below in Figure 17 as the orange-colored node. Many secondary bridged relationships can be seen across 2nd and 3rd degree corporate entities formed by this central intermediary. What is even more interesting is that all corporate entities which are established by this intermediary are all offshore entities which do not originate in its same jurisdiction.

In conclusion, moving up to the super community shows a much different picture and provides insight that the intermediary of the shell companies of the original person being investigated is a part of a much riskier network of individuals and corporate entities. The community shown in Figure 16 is highlighted for perspective in the broader super community shown in Figure 17. This context is not provided by the ICIJ Power Player example, but is discovered by using the community-based drill path outlined in this paper.

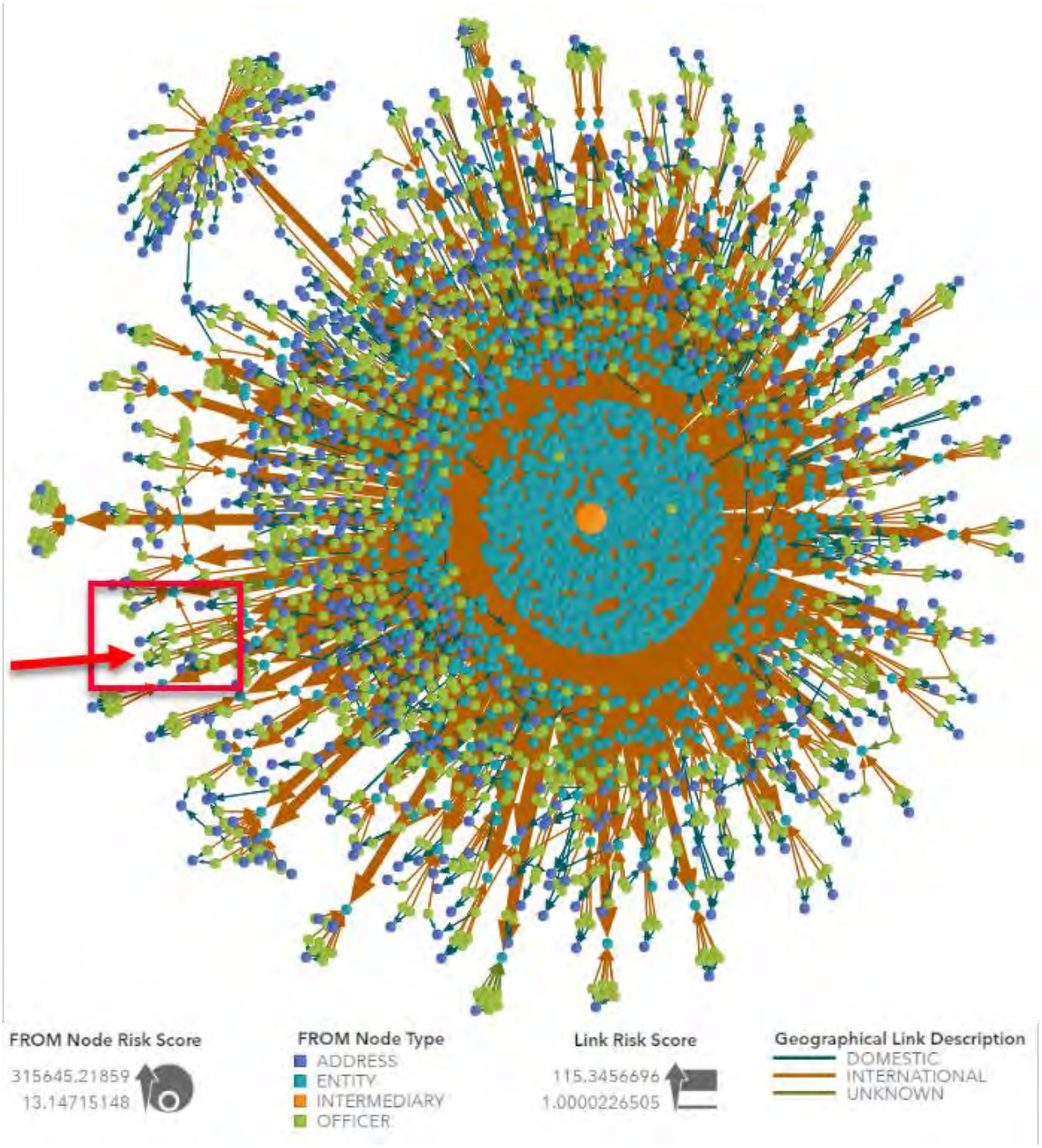


Figure 17: Super community view from original base node in bottom-up analysis.

CONCLUSION

Network analysis is an extensive and complex solution to undertake, but provides a comprehensive process to potentially discover hidden relationships previously unknown to investigators. Interactive analysis of suspicious entities becomes more productive. Automated systems which detect potentially suspicious behavior can become more effective as well. Network analysis provides a new dimension to a risk-hungry organization looking to develop advanced systems to combat financial crime.

REFERENCES

"ICIJ Offshore Leaks Database." The International Consortium of Investigative Journalists. Accessed February 14, 2018. Available at https://offshoreleaks.icij.org/.

"Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)." Creative Commons. Accessed February 14, 2018. Available at https://creativecommons.org/licenses/by-sa/3.0/.

"Open Database License (ODbL) v1.0." Open Data Commons. Accessed February 14, 2018. Available at https://opendatacommons.org/licenses/odbl/1.0/.

"SAS Visual Data Mining and Machine Learning 8.2: The NETWORK Procedure." SAS Institute. Accessed March 6, 2018. Available at http://documentation.sas.com/?docsetId=casmlnetwork&docsetTarget=titlepage.htm&docsetVersion=8.2&locale=en.

"SAS Visual Analytics Customer Documentation Page." SAS Institute. Accessed March 6, 2018. Available at https://support.sas.com/documentation/onlinedoc/va/.

Overton, Stephen. April 28, 2015. "Lasso Your Business Users by Designing Information Pathways to Optimize Standardized Reporting in SAS® Visual Analytics." Proceedings of the SAS Global Forum 2015 Conference. Available at: http://support.sas.com/resources/papers/proceedings15/2960-2015.pdf.

SAS programming code which transforms and builds necessary tables for network analysis described in the Data Management Process section of this paper is available at: https://stephenoverton.net/sas-global-forum-2018/.

ACKNOWLEDGMENTS

The author thanks Falko Schulz for his assistance with the Network Diagram visualization in SAS® Visual Analytics, understanding community analysis with SAS® Viya, and the many review sessions to form ideas on ways to visualize the complex data within the ICIJ Offshore Leaks. The author also thanks Matthew Galati for his assistance on technical challenges with the NETWORK procedure, helping generate centrality statistics for large volumes of data, and for providing guidance on properly using the algorithms within the NETWORK procedure.

RECOMMENDED READING

- "ICIJ Offshore Leaks Power Players" examples available at https://offshoreleaks.icij.org/stories.
- "SAS® Help Center: Examples: NETWORK Procedure" available at http://documentation.sas.com/?docsetId=casmlnetwork&docsetTarget=procnetwork_network_examples.htm&docsetVersion=8.2&locale=en.

CONTACT INFORMATION

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

Stephen Overton
Overton Technologies
(919) 341-9667
soverton@overtontechnologies.com
https://www.stephenoverton.net
https://www.linkedin.com/in/overton/