MWSUG 2016 – Paper AA04

Discover the golden paths, unique sequences and marvelous associations out of your big data using Link Analysis in SAS® Enterprise Miner ™

Delali Agbenyegah, Alliance Data Systems, Columbus, OH

Candice Zhang, Alliance Data Systems, Columbus, OH

ABSTRACT

The need to extract useful information from large amount of data to positively influence business decisions is on the rise especially with the hyper expansion of retail data collection and storage and the advancement in computing capabilities. Many enterprises now have well established databases to capture Omni channel customer transactional behavior at the product or Store Keeping Unit (SKU) level. Crafting a robust analytical solution that utilizes these rich transactional data sources to create customized marketing incentives and product recommendations in a timely fashion to meet the expectations of the sophisticated shopper in our current generation can be daunting. Fortunately, the Link Analysis node in SAS® Enterprise Miner ™ provides a simple but yet powerful analytical tool to extract, analyze, discover and visualize the relationships or associations (links) and sequences between items in a transactional data set up and develop item-cluster induced segmentation of customers as well as next-best offer recommendations. In this paper, we discuss the basic elements of Link Analysis from a statistical perspective and provide a real life example that leverages Link Analysis within SAS Enterprise Miner to discover amazing transactional paths, sequences and links.

INTRODUCTION

A financial fraud investigator may be interested in exploring the relationship between the financial transactions of suspicious customers, the BNI may want to analyze the social network of individuals identified as terrorists to conduct further investigation or a medical doctor may be interested in understanding the association between different medical treatments on patients and their corresponding results. Link analysis is a data mining technique designed to address questions of this sort. It is a subset of network analysis technique used in identifying, exploring and visualizing connections between objects. These objects may include organizations, people, transactions and machines.

Powered by the technological advancement of our age, many retailers now capture every bit of transaction and interaction that customers have with them. Discovering and understanding these relationships and connections within this big data can provide valuable insights to help business decision making.

The Link Analysis node in SAS Enterprise Miner provides a very powerful but easy to use platform for conducting link analysis and visualizing the results in a succinct manner. In this paper, we first explain the fundamentals of link analysis and describe how to conduct a complete link analysis project using SAS Enterprise Miner with real life example in a retail transactional data set up. The paper concludes by highlighting other applications of the Link Analysis node in SAS Enterprise Miner.

FUNDAMENTALS OF LINK ANALYSIS

ASSOCIATION DISCOVERY

Association discovery is the identification of things that happen together. In a transactional set up, association discovery may refer to the identification of products or items that occur together in a particular

transaction. This technique is popularly known as Market Basket Analysis and is very useful in discovering hidden patterns in big transactional data that provides key insights in business decision making. Association discovery rules are typically based on the proportion of times an event occur alone and in combination of other events within a dataset. These rules are simply if/then statements that help reveal relationships between seemingly unrelated occurrences in a database. The rules are stated as 'if item A is part of an event (*antecedent*), then item B is also part of the event (*consequent*) X% of the time'. This is denoted as A->B. A hypothetical association rule could be 'customers who buy suits also buy ties 38% of the time' or 'Patients who undergo vaginal ultrasound and surgical pathology also undergo legally induced abortion 28% of the time'.

There are five major statistical measures used in evaluating association rules as described below:

- **Support**-represents how frequently an item-set occurs in the database
- **Confidence**-the proportion of time a consequent appears given that the antecedent has occurred
- **Expected Confidence**-the number of consequent transactions divided by the total number of transactions
- **Lift**- measures the strength of the association rule. It is defined as the confidence divided the expected confidence.
- **Conviction**- measures the degree of implication of an association rule. It is the ratio of the expected frequency that item set A occurs without B (that is the frequency that the rule makes an incorrect prediction if A and B are independent) divided by the observed frequency of making incorrect predictions

Let A and B represent item sets and T a set of transactions in a given database. Define an association rule A->B as the occurrence of item set A (antecedent) leads to the occurrence of item set B (consequent).

Then the following equation holds:

$$\text{Support}\ (\text{A}{-}{>}\ \text{B}) = \sup(\text{A}, \text{B}) = \text{P}(\text{A n B}) \quad (1)$$

$$\text{Confidence}\ (\text{A}{-}{>}\ \text{B}) = \text{conf}(\text{A}, \text{B}) = \frac{\sup(A,B)}{\sup(A)} = P(B|A) \quad (2)$$

$$\text{Lift}\ (\text{A}{-}{>}\ \text{B}) = \frac{\sup(A,B)}{\sup(A)\sup(B)} = \frac{P(B|A)}{P(B)} \quad (3)$$

$$\text{Conviction}(\text{A}{-}{>}\ \text{B}) = \text{conv}(\text{A}, \text{B}) = \frac{\sup(A)\sup(B')}{\sup(A,B')} = \frac{P(A)(1-P(B))}{P(A)-P(A,B)} \quad (4)$$

It is important to keep in mind that association rules do not imply causation. Additionally in association analysis, only the presence of an item in the transaction is relevant and the number of times an individual buys that item does not matter.

SEQUENCE DISCOVERY

Sequence discovery takes into consideration the ordering of the occurrences in association analysis. In sequence analysis, the rules imply a timing element. The sequence rule A->>B implies that event A occurs before event B. In this situation, event A becomes the *precedent* and event B is the *consequent.* A hypothetical sequence rule could be '38% of customers who buy Digital TVs also buy home theater systems in the next month' or ' 59% of patients who undergo pulmonary bronchospasm evaluation also undergo asthma treatment in the next six months'. The concepts of support, confidence and lift described above under association discovery are also used to evaluate sequence rules.

CENTRALITY MEASURES AND LINK BASED CLUSTERING

One of the powerful advantages of link analysis is to convert association and/or sequence rules into network graphs. Various measures of centrality are then computed and clusters could be developed using these centrality measures. The centrality of a node in a network is a measure of its structural importance. Below we described briefly some centrality measures that are leveraged within link analysis to detect link-based clusters.

- **Degree Centrality:** This refers to the number of ties a node has to other nodes. Actors that have more ties have multiple ways to reach their goals and hence are more advantaged. In a directed graph, the out degree centrality of a node is the number of links that start at this node and connects to another node whiles in-degree is the number of links that start from other nodes and connect to this node. In-degree or out-degree centrality is substitutable in an undirected graph
- **Closeness Centrality:** Closeness centrality is the measure of the degree to which an individual node is near all other nodes in a network. Mathematically, it is the inverse of the sum of the shortest distances between each node.
- **Betweenness Centrality:** Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. It can be thought of as a measure of how quickly information can flow among the nodes in a network.
- **Eigenvector Centrality:** Eigenvector centrality tries to generalize degree centrality by incorporating the importance of neighbors(or incoming neighbors in a directed graph).In real world, having more friends by itself does not guarantee your importance but having more important friends does show more importance. That is exactly what eigenvector centrality tries to solve for. The eigenvector centrality of a node therefore is proportional to the sum of centrality scores of all nodes that are connected to it
- **Influence centrality:** Influence centrality is the generalization of degree centrality that takes into account the link and node weights of adjacent nodes as well as the link weights of nodes that are adjacent to the adjacent nodes.

- **Clustering Coefficient:** It is the measure of the degree to which nodes in a graph tends to cluster together. For a particular node, the clustering coefficient is the ratio of number of links between the nodes within its neighborhood and the number of links that could possibly exist between them.

LINK ANALYSIS IN ENTERPRISE MINER

In SAS Enterprise Miner, link analysis is conducted in a transactional data set up using the link analysis node in the following four major steps:

1) The node first conducts association /sequence rule discovery through computing confidence, support, expected confidence and lift using equations 1, 2 and 3 stated above.
2) The link analysis node then transforms the rules into a network graph data in the form of nodes and links where the support of each item becomes the node weight and the strength of the association (confidence of the rule) becomes the link weight. The two-item sequence rules are transformed into a directed graph data and the association rules into an undirected graph data
3) The node then calculates several centrality measures and detects item clusters from the link graph
4) Finally, the transactional data is scored. There are two score properties under the link analysis node. The node can either produce a next-best-offer list using the association /sequence rules or produce customer segmentation information for scoring using the item clusters.

   If offer recommendation option is chosen, the link analysis node uses weighted confidence to construct the next-best offer list. For each association or sequence rule, the corresponding confidence and support are used to calculate the weighted confidence. The node then looks at each new customer's basket and check which items are already in that particular customer's basket. The candidate items that are already existing are taken out and the remaining candidate items based on the rules are then recommended with their corresponding weighted confidence numbers.

   If segmentation option is chosen, the link analysis node uses item intensity to provide customer segmentation information to help identify different types of customers based on their transaction history. Item clusters are detected from the link graphs such that the links within the item cluster's subgraphs are more densely connected than the links between the item clusters.

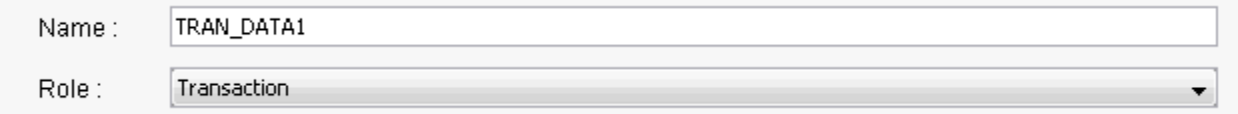
APPLICATION ON TRANSACTIONAL DATA SET UP

- **ASSOCIATION DISCOVERY ANALYSIS**

A modified real data for an apparel retailer is used below to illustrate a full link analysis project. Initial data cleaning and preparation is done within regular Base SAS and the cleaned data brought into the SAS Enterprise Miner's data source and set the data role as transaction as displayed below.

customer_id	item_purchased	number_of_item
1	PENCIL/STRAIGHT	1
2	FLATS	2
2	FULL	2
2	KNIT TOPS	1
2	PENCIL/STRAIGHT	1
2	SHORTS	1
2	SWTR PULLOVER	1
2	WOVEN BLOUSE	2
2	WOVEN DRESS	3
2	WOVEN SHIRTING	1
3	DENIM	2
3	KNIT TEE	1
3	SWTR PULLOVER	1

Display 1. Snapshot of Link Analysis dataset



Display 2. Data Role assignment

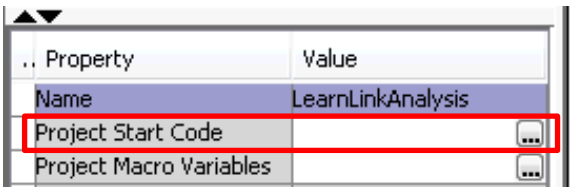
The following steps are used to complete the Link Analysis. We will first configure the settings to run an association discovery and view results and later modify the settings to allow for a sequence discovery.

Step1. Assign the variable roles within the Input Data node.
In the input dataset node, right click each variable to assign the variable roles. In our data set, we set customer_id as the ID variable, item_purchased as the target variable and the number_of_item as the frequency variable. If sequence analysis is needed, an additional time variable is required and is assigned the sequence role.

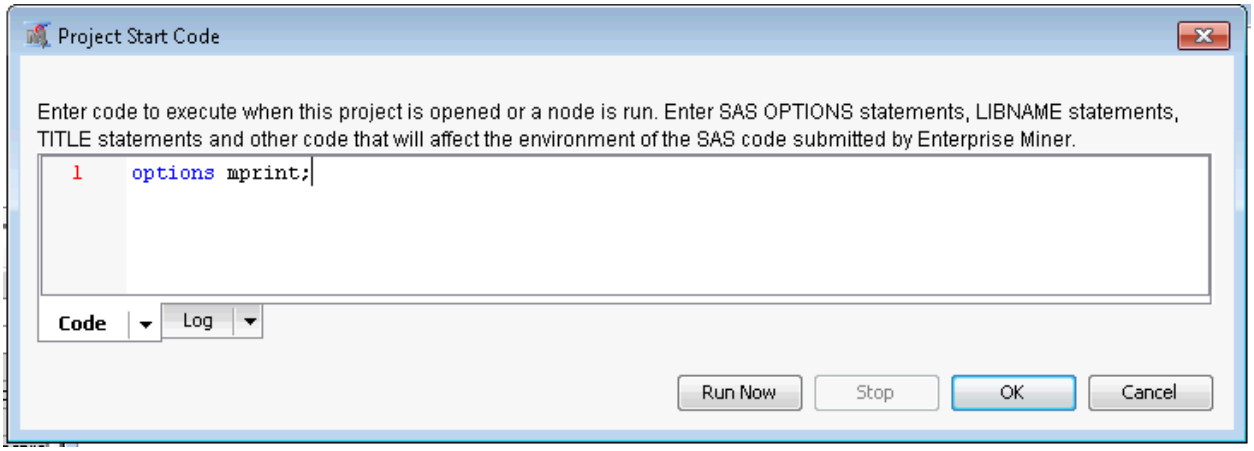
Name	Role	Level
customer_id	**ID**	**Interval**
item_purchased	**Target**	**Nominal**
number_of_item	**Frequency**	**Interval**

Display 3.Variable Role settings

Step2. Set up the link analysis node and modify settings.
To gain visibility to all the outputs and codes needed to complete the link analysis, remember to include options mprint in the project start code editor window as shown below.

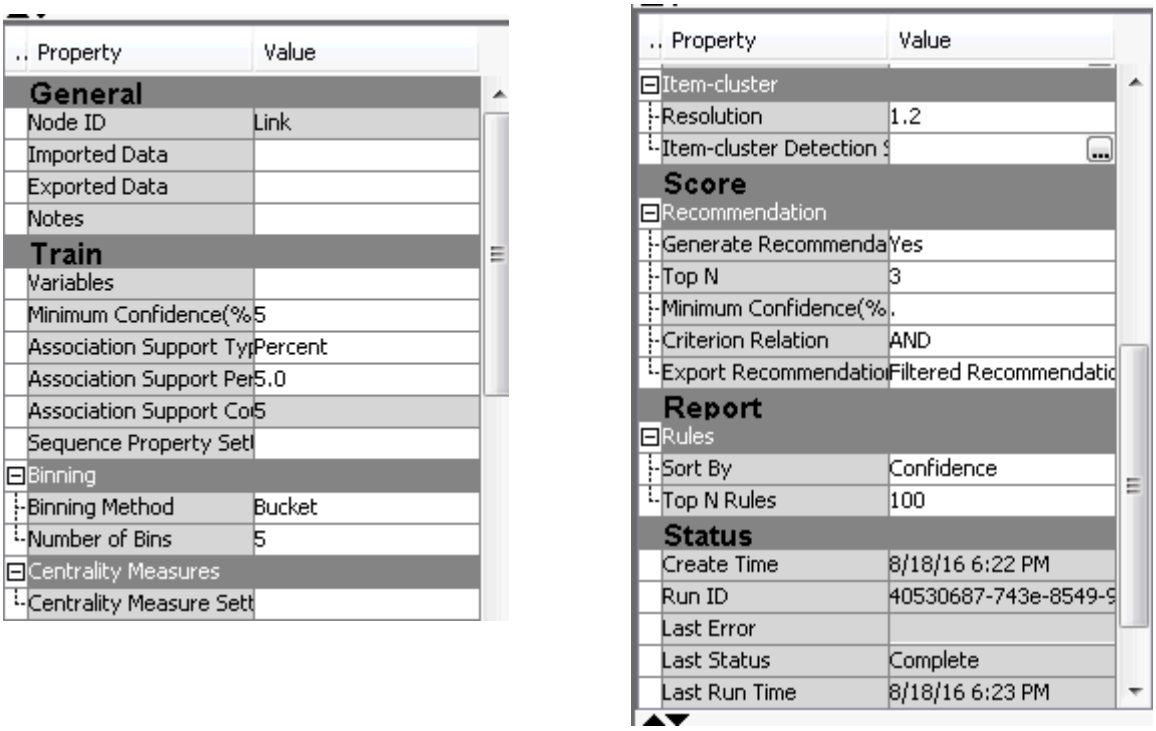


Display 4. Invoking the Project Start Code



Display 5. Project Start Code window

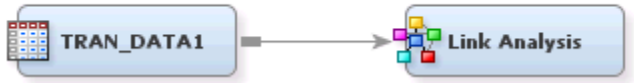
Next, drag the link analysis node to the project diagram window and modify the settings. Users need to set the training, scoring and reporting parameters within the Link Analysis node as shown in our example below.



Display 6.Link Analysis Node Property settings

In our example, we set both our minimum confidence and association support to 5% in the Train section. This means we consider the rule A->B only if P (B|A) is at least 5% and we examine this rule if A and B occurs together at least 5% of the time. Users may also set the association support using counts. The remaining properties are set to the default

Step3. Once the settings are completed, connect the Input data node with the link analysis node and run the project



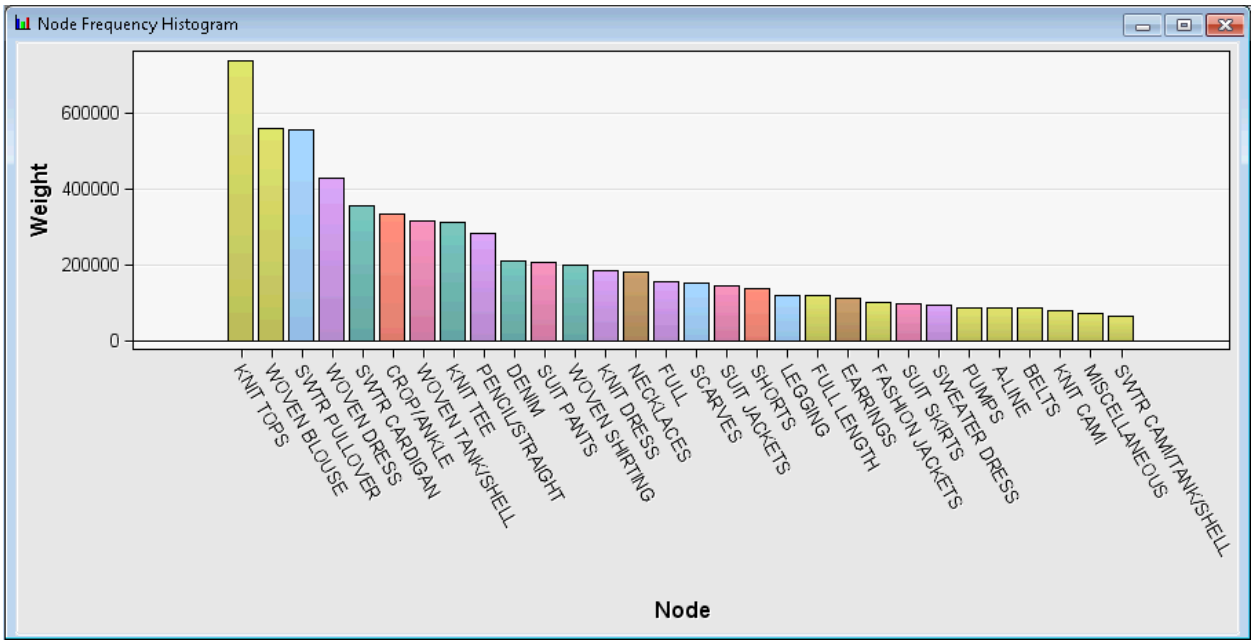
Display 7. Link Analysis process diagram

Step4. View results
After the run completes successfully, we can now view our results and generate valuable insights. The results view generates several outputs to allow us to explore the data, discover the associations and view them graphically in a network graph, view multiple centrality measures as well as show scoring results.

- **EXPLORATORY ANALYSIS**

A. Item Distribution-Node Weight

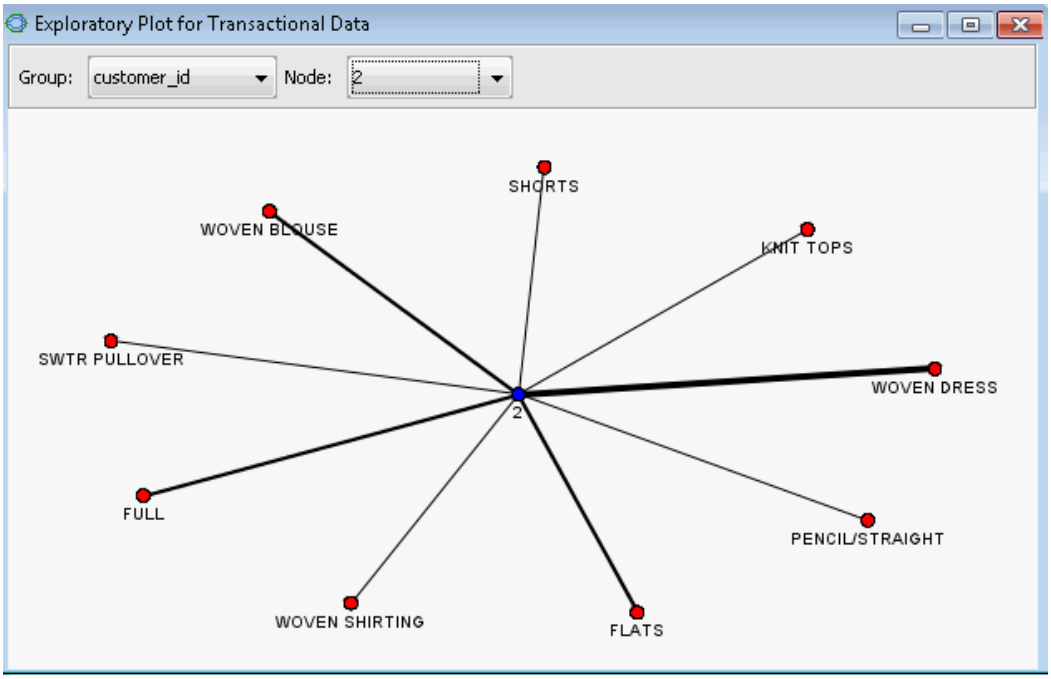
The node weight of an item is the frequency of that item. It simply illustrates which items are most popular. For this retailer, more than 700,000 customers bought their knit top products. Its woven blouses and pull over sweaters are also very well sold. The results are shown in display 8 below.



Display8. Item purchase distribution

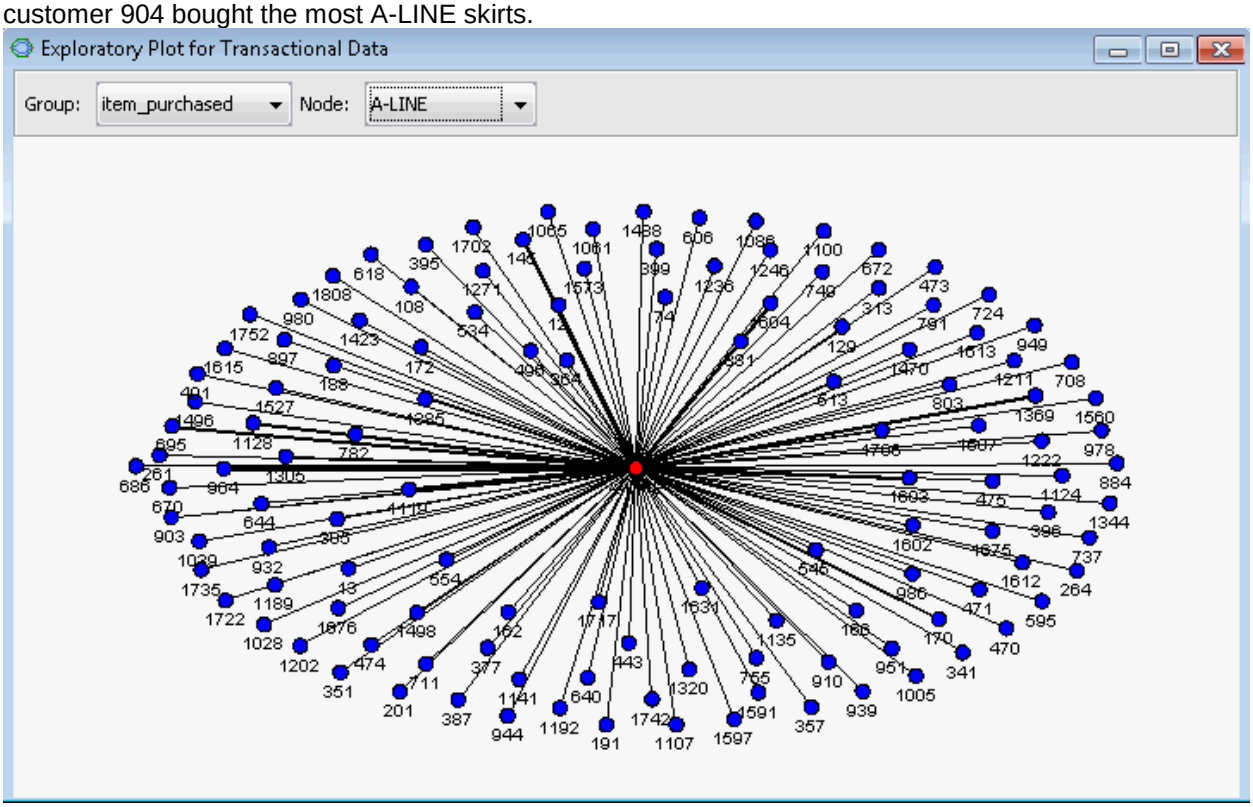
B. Item-customer relations

SAS Enterprise Miner provides a very powerful visualization on item-customer relations. Users can choose to see which customers have bought a specific product or what products a specific customer has bought. For example, the items bought by customer 2 is visualized below and from the thickness of the link, we can say her most frequently purchased item at the brand is woven dress.



Display 9. Plot of Items purchased by a specific customer 2

As mentioned, users can also see which customers bought a certain product the most. It is easy to see that customer 904 bought the most A-LINE skirts.



Display 10. How customers bought A-LINE skirt

- **ASSOCIATION DISCOVERY**

In the results window, users can find view a table of association rules. Remember in the link analysis node settings, we set to see the top 100 association rules ranked by confidence. Below is a screenshot of part

of the rule table. We will take rule 12 as example to illustrate how to read this table. The expected confidence of rule 12 is 9.73%, indicating that suit jackets will be bought 9.73% of the time and a support of 3.56% indicates that suit jackets and suit pants are purchased together 3.56% of the time. When a customer already bought suit jackets, she has a 52.28% of chance to also buy a pair of suit pants as shown by the confidence, which is 5.37 times the chance of buying suit parts within the entire population, shown by the lift value of 5.37.This information can be leveraged when targeting customers for a suit campaign.

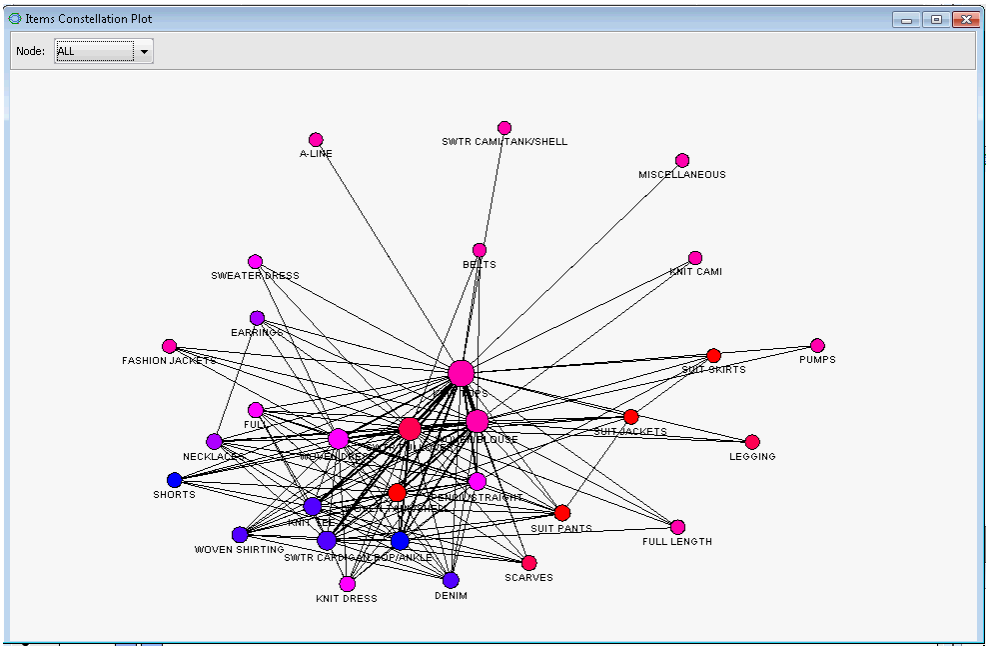
Table: Rule Statistics by Rule ID

Rule ID	Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule	Left Hand of Rule	Recommended Items
1	2	34.97	59.89	1.76	1.71	37139	SWTR CAMI/TANK/SHELL ==> KNIT TOPS	SWTR CAMI/TANK/SHELL	KNIT TOPS
2	2	34.97	58.26	1.99	1.67	42119	MISCELLANEOUS ==> KNIT TOPS	MISCELLANEOUS	KNIT TOPS
3	2	34.97	57.98	8.61	1.66	181869	WOVEN TANK/SHELL ==> KNIT TOPS	WOVEN TANK/SHELL	KNIT TOPS
4	2	34.97	56.05	2.26	1.60	47663	A-LINE ==> KNIT TOPS	A-LINE	KNIT TOPS
5	2	34.97	55.72	2.61	1.59	55078	FASHION JACKETS ==> KNIT TOPS	FASHION JACKETS	KNIT TOPS
6	2	34.97	54.03	2.00	1.54	42237	KNIT CAMI ==> KNIT TOPS	KNIT CAMI	KNIT TOPS
7	2	34.97	52.68	2.09	1.51	44226	BELTS ==> KNIT TOPS	BELTS	KNIT TOPS
8	2	34.97	52.42	2.41	1.50	50927	SUIT SKIRTS ==> KNIT TOPS	SUIT SKIRTS	KNIT TOPS
9	2	34.97	52.36	2.76	1.50	58286	EARRINGS ==> KNIT TOPS	EARRINGS	KNIT TOPS
10	2	34.97	52.36	7.73	1.50	163292	KNIT TEE ==> KNIT TOPS	KNIT TEE	KNIT TOPS
11	2	34.97	52.35	6.99	1.50	147746	PENCIL/STRAIGHT ==> KNIT TOPS	PENCIL/STRAIGHT	KNIT TOPS
12	2	9.73	52.28	3.56	5.37	75124	SUIT JACKETS ==> SUIT PANTS	SUIT JACKETS	SUIT PANTS
13	2	34.97	52.25	13.83	1.49	292108	WOVEN BLOUSE ==> KNIT TOPS	WOVEN BLOUSE	KNIT TOPS
14	2	34.97	52.08	2.92	1.49	61742	FULL LENGTH ==> KNIT TOPS	FULL LENGTH	KNIT TOPS
15	2	34.97	51.96	4.43	1.49	93621	NECKLACES ==> KNIT TOPS	NECKLACES	KNIT TOPS
16	2	34.97	51.49	2.91	1.47	61473	LEGGING ==> KNIT TOPS	LEGGING	KNIT TOPS
17	2	34.97	51.45	8.07	1.47	170466	CROP/ANKLE ==> KNIT TOPS	CROP/ANKLE	KNIT TOPS
18	2	34.97	50.64	4.93	1.45	104131	SUIT PANTS ==> KNIT TOPS	SUIT PANTS	KNIT TOPS
19	2	34.97	50.30	3.21	1.44	67889	SHORTS ==> KNIT TOPS	SHORTS	KNIT TOPS
20	2	34.97	50.26	2.03	1.44	42799	PUMPS ==> KNIT TOPS	PUMPS	KNIT TOPS

Display 11. Table of association rules

- **ITEMS CONSTELLATION PLOTS**

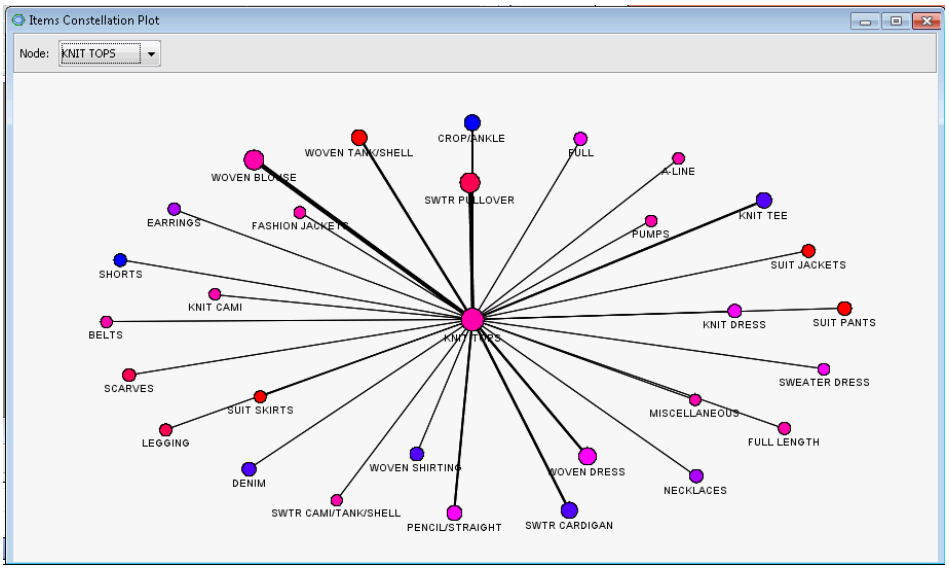
The Items Constellation Plot give association graphs among items. Users can choose to see the constellation plot of all items or a single item with its neighbors. The color of nodes represent item clusters, which will be explained in more details in the next session. The size of nodes represent the support of the item and the thickness of the link represents association strength between items. The visualization of the associations among all items is shown is display 12



Display 12. Items constellation plot

If a certain item is of interest, users can choose the item in “Node” list. For example, if we want to see what items are usually purchased with knit tops, we just need to select “Knit Tops” in the node list and the plots

will show as display 13. It seems like knit tops are most often purchased with woven blouses, woven tanks and pullover sweaters.

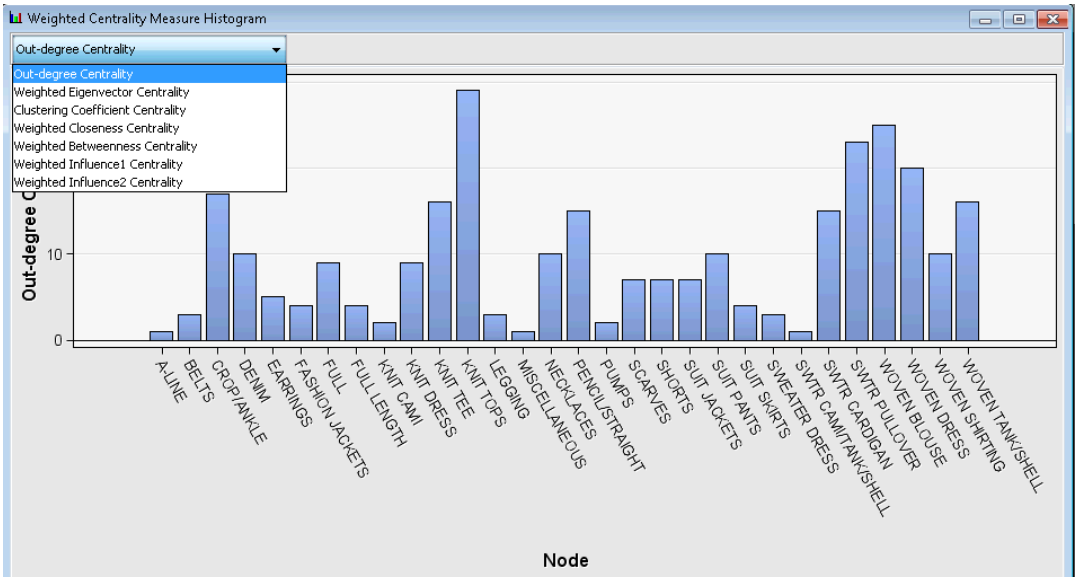


Display13.Items constellation plot for KNIT TOPS

- **ITEM CLUSTERING**

A. Centrality measures

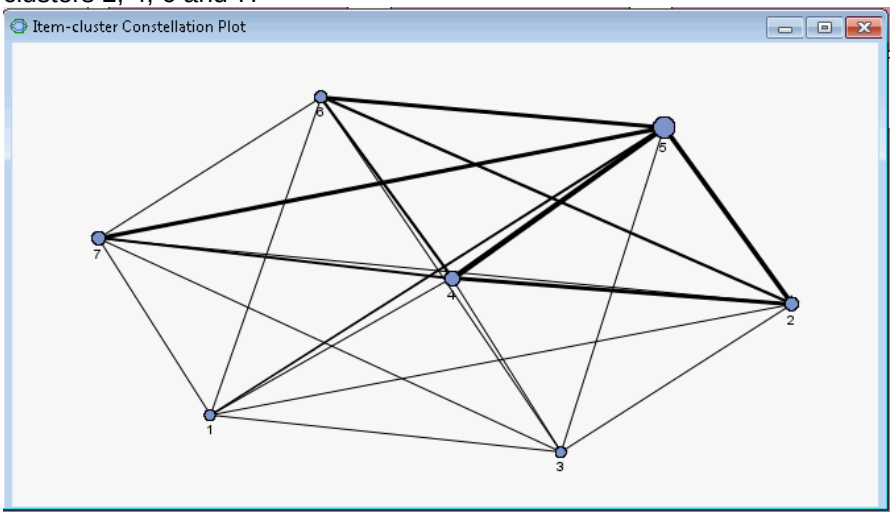
Centrality measures quantify how important a node is in the association community. There are several measures of centrality described earlier in this paper, each with a specific focus. In the Weighted Centrality Measure Histogram menu, users can choose the measure of interest. In the example below, we show the out-degree centrality measure to see which items have most connections with other items. Not surprising that knit tops have most connections with other items since it's also the most popular item. A-line products seem to have limited connections with other items. Actually, from the item constellation plot, we also know that A-line only appears together with knit tops. This means customers who bought A-line have little preference of other products except knit tops.



Display14.Weighted Centrality Measure Histogram

B. Item clusters

One of the applications of link analysis is to conduct item clustering. There are 109 items in the transactional data we used in our example. 30 of them meet the confidence and support threshold. The association among these 30 items are mined in the link analysis and they are segmented into 7 clusters. In the plot below, the association strength between clusters are represented by the thickness of the link and the cluster sizes are represented by the size of respective node. Cluster 5 is the largest and has a strong relation with clusters 2, 4, 6 and 7.



Display15.Item Cluster Constellation Plot

The table below also shows the list of items in each cluster and the results makes sense intuitively. Generally, suits products are grouped together and jewelry products also appear in the same cluster.

Cluster 1	Crop/Ankle	Shorts								
Cluster 2	SWTR Cardigan	Knit Tee	Denim	Woven Shirt						
Cluster 3	Necklaces	Earrings								
Cluster 4	Woven Dress	Pencil/Straight	Knit Dress	Full	Sweater Dress					
Cluster 5	Knit Tops	Woven Blouse	Full Length	Fashion Jackets	Pumps	A-Line	Belts	Knit Cami	Miscellaneous	SWTR Cami/Tank/Shell
Cluster 6	SWTR Pullover	Scarves	Legging							
Cluster 7	Woven Tank/Shell	Suit Pants	Suit jackets	Suit Skirts						

Display 16. List of Items within each cluster.

- **SCORING RESULTS**

A. Customer segmentation

When user sets 'Generate Recommendation' to be 'No' under the link analysis' score settings, SAS Enterprise Miner will automatically provide a scoring code to conduct customer segmentation based on item intensity. The display below shows an example output where each customer is classified into a cluster (cluster number) based on the items they purchased.

customer_id	final_seg
100	2
101	5
102	5
103	5
104	0
105	4

Display 17. Customer segmentation scoring results.

B. Next best offer recommendation

When user sets 'Generate Recommendation' to be 'Yes' under the link analysis' score settings, SAS Enterprise Miner will automatically provide the top N next best offers for each customer and report their corresponding weighted confidence numbers. In the example below, customer 1 seems to have not purchased crop/ankle pants, full length pants and denim products at the brand yet. Based on the association discovery among the overall customer base and what this customer has already bought, these three products are most likely to be purchased by customer 1.

Filtered Recommendation Table

ID Variable	Next Best Offer	Confidence
1	CROP/ANKLE	24.53984
1	FULL	16.65893
1	DENIM	14.27868
2	CROP/ANKLE	24.25306
2	BELTS	6.517416
2	A-LINE	6.449426
3	CROP/ANKLE	25.01571
3	EARRINGS	8.298011
3	BELTS	6.668575

Display 18. Next best offer recommendation table

- **SEQUENTIAL DISCOVERY ANALYSIS**

Link analysis node can also be used for sequential discovery when there is a transaction date in the data. Sample data could look like display 19 below.

customer_id	item_purchased	tran_date
1	PENCIL/STRAIGHT	27JUN2016
2	FLATS	07SEP2015
2	FULL	07SEP2015

Display 19. Snapshot of Sequential discovery dataset

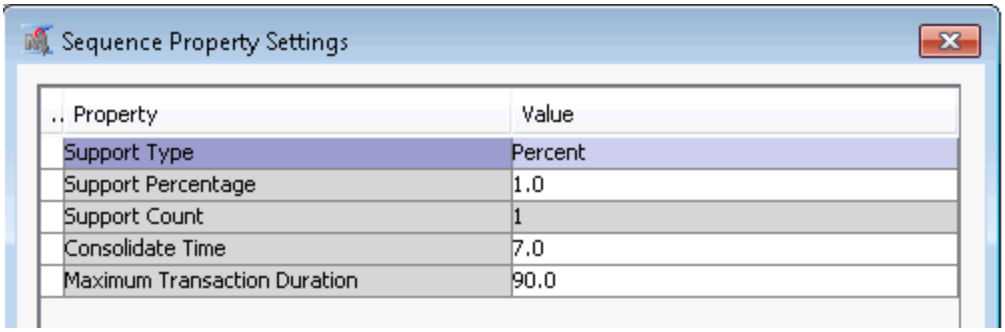
During the variable set up, we need to set tran_date as a role of "Sequence".

Name	Role	Level
customer_id	**ID**	**Interval**
item_purchased	**Target**	**Nominal**
tran_date	**Sequence**	**Interval**

Display 20. Data role assignment under sequential discovery

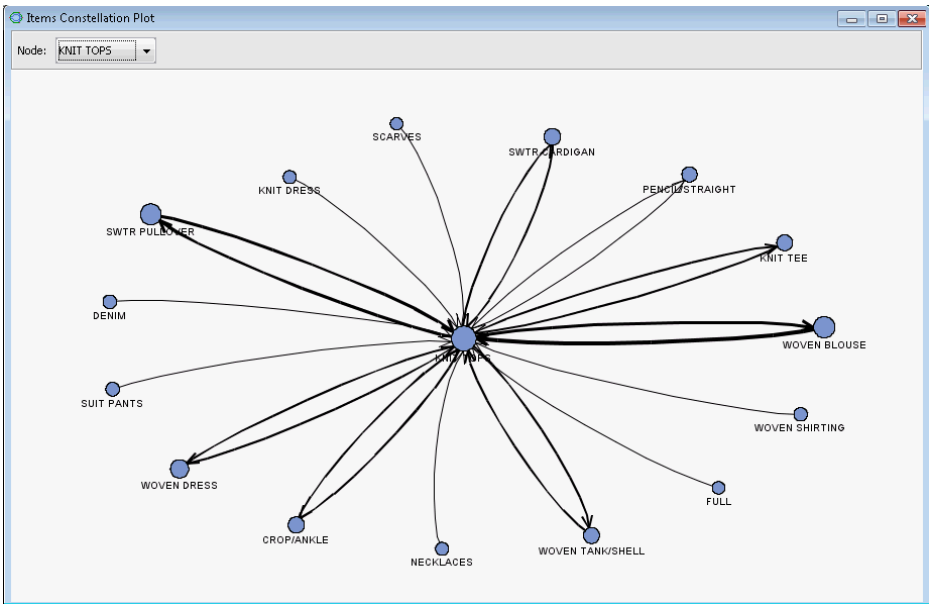
The link analysis node setting for association discovery is very similar to the settings for sequential discovery except for the data role assignment step where we set tran_date as a sequence variable. Additionally, there is a "Sequence Property Setting" in the node property that has to be set. Minimum count of support to claim item association can be set as percent or count threshold. In our example below, we kept the default setting. A sequence need to appear at least 1% of the time to be claimed as a sequential rule. Consolidate Time specifies whether repeated purchase of an item within a time frame can be considered as one purchase. In the example we set it to 7 days, meaning items purchased within a week can be consolidated. Maximum Transaction Duration specifies the maximum transaction window length. Setting this at 90 in our example means any successive purchases with a gap of more than 90 days will not be considered as a sequential rule. By default, there's no restriction on consolidate time and maximum transaction duration. However, adding these constrains can often greatly improve efficiency and make the

results more meaningful. We strongly recommend adjusting these parameters based on business needs and the project goal.



Display 21.Sequence Property Settings

In sequential discovery, relations between items are mined similarly as in association analysis, except that the purchasing order is important in determining the sequential rule. Consequently, the relationships between items are represented by directed links. The arrow illustrates the order of occurrence and it points from antecedent to consequent. In the example below in display 22, we see that customers may purchase woven blouse after knit top or purchase knit top after woven blouse. However, we rarely see customers buy suit pants after buying a knit top.



Display 22.Item Constellation Plot for sequential discovery.

Next best offer will also be generated for sequential analysis if the ‘Generate Recommendation’ is set to be “Yes”. However, the customer segmentation and item clustering is not conducted under link analysis since it makes little sense.

Filtered Recommendation Table

ID Variable	Next Best Offer	Confidence
1	CROP/ANKLE	13.4379
1	KNIT DRESS	7.686393
1	DENIM	7.504278
2	CROP/ANKLE	12.47532
2	DENIM	7.151911
2	KNIT DRESS	6.284645
3	CROP/ANKLE	13.82057
3	KNIT DRESS	5.853047
3	FULL	5.440873

Display 18. Next best offer recommendation table for a sequential discovery

CONCLUSION

Link analysis is a powerful data mining technique to discover useful associations and sequences hidden in large data sets. We have illustrated how one can easily conduct this analysis within SAS Enterprise Miner using the Link Analysis Node. SAS Enterprise Miner also has the Association Node that can be leveraged to undertake both association and sequential analysis which can also generates two or more item association or sequence rules.

Link analysis can be applied in other industries outside retail to discover hidden patterns in any database. The police departments can use it for criminal investigations and crime detection, financial institutions can leverage it for fraud detection, the FBI can conduct social network analysis the track the connections of terrorists and other suspicious people in social media and many organizations can utilize link analysis to evaluate their communication and collaboration as well as understand relationships and identify communities across geographies and divisions. SAS Enterprise Miner has made all these possible in a matter of clicks!

REFERENCES

- Liu, Y., Lee T., Zhang R., Dean J. (2014) Link Analysis using SAS Enterprise Miner, Cary, NC: SAS Institute Inc.
- Xinli, B. (2007) Mining transaction/Order Data Using SAS Enterprise Miner Association Node, SAS Global Forum 2007 Paper 132-2007,Orlando,FL
- Michael, L. (2014). Visualizing Big Data: Social Network Analysis, Digital Research Conference, San Antonio, Texas.

ACKNOWLEDGMENTS

The authors will like to thank Yin Chen and the entire Predictive Analytics team at Alliance Data Card Services for their support.

CONTACT INFORMATION

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

Delali Agbenyegah
Alliance Data Systems
3100 Easton Square Place
Columbus, OH, 43219
delali.agbenyegah@alliancedata.com

Candice Zhang
Alliance Data Systems
3100 Easton Square Place
Columbus, OH, 43219
Candice.Zhang@alliancedata.com