

*IBM SPSS Modeler Social Network
Analysis 17.1 User Guide*

IBM

Note

Before using this information and the product it supports, read the information in "Notices" on page 19.

Product Information

This edition applies to version 17, release 1, modification 0 of IBM(r) SPSS(r) Modeler and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Preface	v	Partitioning into groups	10
About IBM Business Analytics	v	Describing groups and group members	10
Technical Support	v	Group Analysis node	11
Chapter 1. IBM SPSS Modeler Social Network Analysis.	1	Specifying data for group analysis.	11
About IBM SPSS Modeler Social Network Analysis	1	Setting build options for group analysis	12
About social network analysis	1	Viewing group analysis statistics	12
Displaying networks.	2	Output for group analysis	13
Describing networks.	3	Chapter 3. Diffusion analysis	15
IBM SPSS Modeler Social Network Analysis nodes	5	Diffusion analysis overview	15
Node tabs	5	Diffusion analysis example	15
Previewing output	5	Diffusion Analysis node	16
Analyzing data	6	Specifying data for diffusion analysis.	16
Applications	6	Setting build options for diffusion analysis	17
Data structure	7	Viewing diffusion analysis statistics	17
Scripting properties	7	Output for diffusion analysis	18
Chapter 2. Group analysis	9	Notices	19
Group analysis overview	9	Trademarks	20
Determining social similarity	9	Index	25

Preface

IBM® SPSS® Modeler Social Network Analysis processes information about relationships between people into fields describing an individual's role in a social network, allowing social information to be included in predictive models. This manual describes the use of the IBM SPSS Modeler Social Network Analysis nodes in the IBM SPSS Modeler environment, enabling you to include the nodes in your streams. When the node output is combined with fields representing measurements on individuals, a more complete profile of the individuals results.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises – able to direct and automate decisions to meet business goals and achieve measurable competitive advantage. For further information or to reach a representative visit <http://www.ibm.com/spss>.

Technical Support

Technical support is available to maintenance customers. Customers may contact Technical Support for assistance in using IBM Corp. products or for installation help for one of the supported hardware environments. To reach Technical Support, see the IBM Corp. web site at <http://www.ibm.com/support>. Be prepared to identify yourself, your organization, and your support agreement when requesting assistance.

Chapter 1. IBM SPSS Modeler Social Network Analysis

About IBM SPSS Modeler Social Network Analysis

Many approaches to modeling behavior focus on the individual. They use a variety of data about individuals to generate a model that uses the key indicators of the behavior to predict it. If any individual has values for the key indicators that are associated with the occurrence of the behavior, that individual can be targeted for special attention designed to prevent the behavior.

Consider approaches to modeling churn, in which a customer terminates his or her relationship with a company. The cost of retaining customers is significantly lower than the cost of replacing them, making the ability to identify customers at risk of churning vital. An analyst often uses a number of Key Performance Indicators to describe customers, including demographic information and recent call patterns for each individual customer. Predictive models based on these fields use changes in customer call patterns that are consistent with call patterns of customers who have churned in the past to identify people having an increased churn risk. Customers identified as being at risk receive additional customer service or service options in an effort to retain them.

These methods overlook social information that may significantly affect the behavior of a customer. Information about a company and about what other people are doing flows across the relationships to impact people. As a result, relationships with other people allow those people to influence a person's decisions and actions. Analyses that include only individual measures are omitting important factors having predictive capabilities.

IBM SPSS Modeler Social Network Analysis addresses this problem by processing relationship information into additional fields that can be included in models. These derived key performance indicators measure social characteristics for individuals. Combining these social properties with individual-based measures provides a better overview of individuals and consequently can improve the predictive accuracy of your models.

IBM SPSS Modeler Social Network Analysis consists of two primary components:

- IBM SPSS Modeler Social Network Analysis nodes added to the IBM SPSS Modeler environment that enable the inclusion of social analytic techniques in streams.
- IBM SPSS Modeler Server Social Network Analysis, which adds processing of the node specifications to IBM SPSS Modeler Server. IBM SPSS Modeler Server Social Network Analysis efficiently processes massive amounts of network data, which may include millions of individuals and relationships, into a relatively small number of fields for further analysis.

For example, IBM SPSS Modeler Social Network Analysis identifies the individuals in a network that are most affected by specific people churning. Furthermore, you can discover groups of individuals in a network that are at an increased risk of churn. By incorporating Key Performance Indicators for these effects in your models, you can improve their overall performance.

About social network analysis

A social network consists of a set of individuals and the relationships between them. Social network analysis examines these relationships to describe individuals and groups as parts of a social structure. Individuals interact with each other and these interaction patterns provide insight into the individuals involved. Relationships enable information to flow across a network, enabling one individual to influence another. The importance of the relationship information sets social network analysis apart from other approaches. Instead of focusing on each individual separately, the unit of study is a dyad consisting of two individuals and their relationships.

Relationships in a network can be classified as either directional or nondirectional. In a *directional relationship*, one individual is identified as the initiator, or source, of the relationship and the other is identified as the receiver, or destination. For example, making a phone call is a directional relationship in which one person calls another. In contrast, the roles of source and destination cannot be defined for *nondirectional relationships*. In this case, both parties participate in the relationship equally. Speaking to each other is an example of a nondirectional relationship.

Another property that distinguishes between relationships is whether the relationship is dichotomous or valued. The only information available in a *dichotomous relationship* is whether or not the relationship exists between two individuals. For every dyad in the network, the relationship is either present or absent. A *valued relationship*, on the other hand, includes a weight indicating the strength of the relationship. The weights allow the relationships to be compared to each other.

The "Relationship types" table lists examples for the cross-classification of relationships by direction and scale. In the directional relationships, *Joe* is the source of the relationship and *Mary* is the destination. In the nondirectional relationships, there is no indication of who initiated the relationship. The valued relationships use the length of the conversation as the relationship weight, while the dichotomous relationships either occurred or they did not.

Table 1. Relationship types.

Direction	Scale	Example
Nondirectional	Dichotomous	Joe and Mary spoke to each other
Nondirectional	Valued	Joe and Mary spoke to each other for 20 minutes
Directional	Dichotomous	Joe called Mary
Directional	Valued	Joe called Mary for a 20 minute conversation

For more information about the field of social network analysis, consult one of the comprehensive books in this area ¹.

Displaying networks

A social network is typically illustrated using a *sociogram* ². In this type of visual display, individuals correspond to points, or nodes, in a space. Lines, or edges, connecting the points represent relationships between the individuals. If the relationships are directional, the edges include an arrow to indicate the direction. If the relationships have weights, the labels for the edges denote the values. The following graph displays a network for seven individuals.

1. Wasserman, S., and K. Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

2. Moreno, J. L. 1934. *Who Shall Survive?: Foundations of Sociometry, Group Psychotherapy, and Sociodrama*. Washington, D.C.: Nervous and Mental Disease Publishing Co..

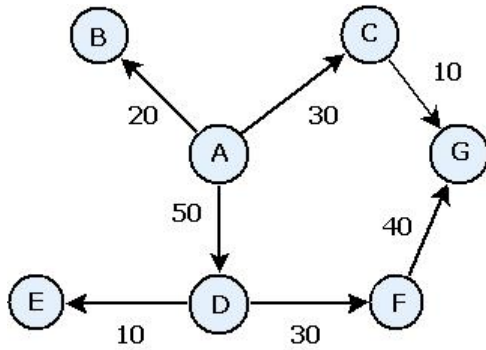


Figure 1. Example social network

Suppose the network represents the phone calls made by individuals with the relationship weights indicating the length of the calls. In this case, Person A called three people, spending the majority of time talking to person D.

This network is much smaller than those encountered in practice. However, the concepts illustrated by simple sociograms generalize to networks of any size and complexity.

Describing networks

Information about networks, groups, and individuals needs to be extracted into descriptive characteristics that allow cross-comparisons and inclusion in predictive models. Networks need to be distilled into a finite set of key performance indicators that can be analyzed. For example, you may want to compare networks or groups of nodes within a network to each other. Alternatively, you may want to compare individuals in the network to each other or identify the most important individuals.

Two measures commonly used to describe social networks are **density** and **degree**. Both statistics reflect connectivity, but the former focuses on the entire network or on network subgroups while the latter characterizes the individuals in the network.

Network density

For any set of nodes in a network, there is a finite number of relationships possible. Each node can serve as the source or the target of a relationship with every other node. Consider a network consisting of the three nodes A, B, and C. The following table lists all possible directed relationships between the nodes.

Table 2. Possible directed relationships for three nodes.

Source	Target
A	B
A	C
B	A
B	C
C	A
C	B

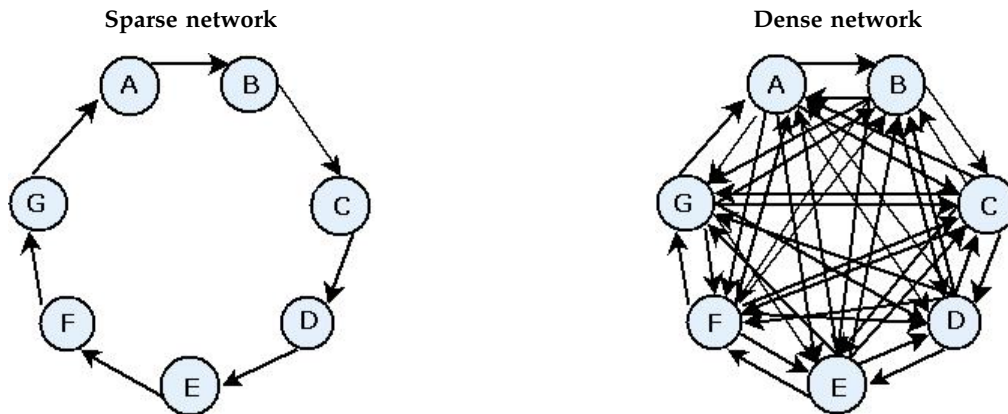
Each node is the source of a relationship with the other two nodes. However, in practice, all possible relationships may not actually be present. Some nodes may not have any direct relationship with other nodes. In addition, some directed relationships may not be reciprocated.

The *density* statistic represents the proportion of possible relationships in the network that are actually present. The value ranges from 0 to 1, with the lower limit corresponding to networks with no

relationships and the upper limit representing networks with all possible relationships. The closer the value is to 1, the more dense is the network and the more cohesive are the nodes in the network.

Information in dense networks can flow more easily than information in sparse networks. The "Sparse and dense networks" table displays two networks consisting of seven nodes. The sparse network includes only seven of the possible 42 relationships between the nodes, yielding a density of 0.17. The dense network, on the other hand, contains all possible relationships and has a density of 1.0.

Table 3. Sparse and dense networks.



In the sparse network, for information to flow from node A to node G, it must pass through five other nodes. In contrast, in the dense network, the information can go directly from node A to node G.

Nodal degree

The important individuals in a network are often the ones involved in the most relationships. These individuals acquire information from a variety of sources and spread that information to a large number of other individuals. In contrast, individuals who participate in a few relationships cannot directly influence a large number of others in the network.

The *degree* for a node, defined as the total number of relationships involving that node, permits comparisons between network participants. Individuals with higher degree values are more active than those with lower values. Degree ignores the direction of the relationships, providing an overall measure of activity for the node.

For directed relationships, you can focus on whether a node is a source or target when tallying the number of relationships. The *indegree* for a node is the number of relationships in which a particular node is the target. Conversely, the *outdegree* is the number of relationships in which a node is the source. The following table lists the degree, indegree, and outdegree values for each node in the "Example social network" figure.

Table 4. Example degree, indegree, and outdegree values.

Node	Degree	Indegree	Outdegree
A	3	0	3
B	1	1	0
C	2	1	1
D	3	1	2
E	1	1	0
F	2	1	1
G	2	2	0



Indegree is often treated as a measure of prestige. Higher indegree values correspond to more relationships ending at that node. In other words, those individuals are contacted by a high number of other individuals. Many other nodes are initiating relationships with the node. Conversely, outdegree is treated as a measure of centrality. Higher values correspond to more relationships originating from that node. Those individuals contact a high number of other individuals.

For the nodes in the example network, the degree values indicate that nodes A and D are the most active while nodes B and E are the least active. The indegree values reveal that node G has the most prestige. Based on the outdegree values, node A is the most central.

IBM SPSS Modeler Social Network Analysis nodes

Along with the many standard nodes delivered with IBM SPSS Modeler, you can also work with IBM SPSS Modeler Social Network Analysis nodes to include the results of social network analysis in your streams. The "IBM SPSS Modeler Social Network Analysis nodes" table describes these nodes, which are stored in the Sources palette.

Table 5. IBM SPSS Modeler Social Network Analysis nodes.

Node	Icon	Description
Group Analysis		The Group Analysis node imports call detail record data from a fixed-field text file, identifies groups of nodes within the network defined by the records, and generates key performance indicators for the groups and individuals in the network. See the topic "Group analysis overview" on page 9 for more information.
Diffusion Analysis		The Diffusion Analysis node imports call detail record data from a fixed-field text file, propagates an effect across the network defined by the records, and generates key performance indicators summarizing the results of the effect on individual nodes. See the topic "Diffusion analysis overview" on page 15 for more information.

Node tabs

The IBM SPSS Modeler Social Network Analysis nodes offer the following tabs for defining and previewing the analysis:

- **Data tab.** Used to identify the file containing the social network information.
- **Build Options tab.** Used to define settings for the analyses.
- **Analysis tab.** Used to view a summary of preliminary output that provides guidance for modifying the input settings on the Data tab to produce optimal results.

In addition, the nodes offer the following tabs common across source nodes in IBM SPSS Modeler:

- **Filter tab.** Used to eliminate or rename the output fields produced by the nodes. This tab offers the same functionality as the Filter node.
- **Types tab.** Used to set measurement levels for the output fields produced by the nodes. This tab offers the same functionality as the Type node.
- **Annotations tab.** Used to rename nodes, supply a custom ToolTip, and store a lengthy annotation.

For more information about common tabs, consult the IBM SPSS Modeler documentation.

Previewing output

Typically, you use the output of the nodes as input to a predictive model. If you want to view the output, you could add a Table node to the stream. However, the amount of data is usually extremely large, making table generation time-consuming. Moreover, the number of rows in the table limits its usefulness.

As an alternative, you can generate a preview table that displays a sample of the output that will be created. The preview shows the generated fields for a limited number of rows. The number of rows is defined in the stream properties. To preview the node output, perform the following steps:

1. Open the node.
2. Specify the data settings on the Data tab.
3. Define the analytical parameters on the Build Options tab.
4. Click **Preview**.

A preview window opens, displaying the results. In addition, previewing the output populates the Analysis tab of the node with a summary overview of the results.

Analyzing data

Determining the analytical settings that produce optimal results is usually an iterative process. You define the settings, run the analysis, and review the results. If the results are not as useful as they could be, you modify the settings and rerun the analysis.

To analyze the input data for the node, perform the following steps:

1. Open the node.
2. Specify the data settings on the Data tab.
3. Define the analytical parameters on the Build Options tab.
4. Click **Analyze Data**.

If the build options indicate summary statistics should be displayed, the Analysis tab shows the results.

If you need to rerun an analysis, click **Clear Analysis** to purge the current results before clicking **Analyze Data**.

Applications

Specific applications in which IBM SPSS Modeler Social Network Analysis may be particularly beneficial include the following:

- **Churn prediction.** Group characteristics can influence churn rates. By focusing on individuals in groups that are at increased risk of churning, you may be able to prevent it. In addition, you can identify individuals that are at risk of churning due to the flow of information from those that have already churned.
- **Leveraging group leaders.** Group leaders are highly influential over other group members. If a group leader can be prevented from churning, the churn rate for the group members may be reduced. Alternatively, attempting to get a group leader from a competitor to churn may increase the churn rate of group members associated with that competitor while reducing the churn rate of group members associated with your company.
- **Marketing.** Group leaders can be used to initiate new goods or service offerings. The influence of the leader may make other group members more likely to purchase the offering. You can use diffusion analysis to identify the individuals most affected by the group leaders and target your marketing to them.

Two demonstration streams are supplied with IBM SPSS Modeler Social Network Analysis to give you examples of how you could include the results of social network analysis in your streams. The data files and sample streams are installed in the *Demos* folder under the product installation directory.

- *DA demo streams.str* gives an example of analyzing data to identify the top 300 individuals who are most likely to churn.

- *GA demo streams.str* gives an example of using key performance indicators (KPIs) to predict churn of both groups and individuals, and also using KPIs to target specific individuals for a marketing campaign.

Data structure

Information about the individuals in the network of interest may be spread across a variety of files, databases, and systems throughout your enterprise. To analyze the network using IBM SPSS Modeler Social Network Analysis, you need to extract the relevant records and fields from your data sources and format them for input to the nodes.

The analytical nodes require call detail records stored in a single, fixed-width text file. Each row of the file corresponds to a relationship, with the data organized in the following columns:

- the identifier for the individual who initiates the relationship.
- the identifier for the individual who is the target of the relationship.
- an optional weight for the relationship.

All data must be numerical, with the identifiers for individuals limited to integers. Inclusion of field names as the first row of data in the file is optional. The "Example call data records" table illustrates this data structure.

Note that header records must also use either numbers or blank values. For example, a header with the value of *EF BB BF* would cause an error.

Table 6. Example call data records.

Source	Destination	Weight
1000	5642	243
2190	8444	831
0299	9419	559

The weight values can correspond to any measure you want to use to represent the importance of the relationship relative to the other relationships in the network. For call data, common weights include the call duration or the call frequency. Note that this is true for Diffusion Analysis, but Group Analysis only supports call frequencies.

If you wish the analysis to focus on a subset of the calling history, you must use that subset when creating the input file. For example, you can limit the analysis to the past several months or to the most recent calls for an individual by including only that data in the input text file.

Scripting properties

Scripting in IBM SPSS Modeler Social Network Analysis is a powerful tool for automating processes in the user interface. Scripts can perform the same types of actions that you perform with a mouse or a keyboard, and you can use them to automate tasks that would be highly repetitive or time consuming to perform manually. For an explanation of using scripting, see the *ScriptingAutomation.pdf* guide available with IBM SPSS Modeler.

Diffusion node properties

The following table lists the scripting properties for the Diffusion node.

Property name	Data type	Property description
input_data_file_name	string	

Property name	Data type	Property description
calling_field	<i>field</i>	
called_field	<i>field</i>	
frequency_weight_field	<i>field</i>	
read_field_names	<i>boolean</i>	
diffusion_list_file_name	<i>string</i>	
spreading_factor	<i>double</i>	Default value = 50. Min = 1 Max = 99.
max_number_iterations	<i>integer</i>	Default value = 100. Min = 1.
accuracy_threshold	<i>double</i>	Default value = 0.01. Min = 0.001.
calculate_statistics	<i>boolean</i>	

Group Analysis node properties

The following table lists the scripting properties for the Group Analysis node.

Property name	Data type	Property description
input_data_file_name	<i>string</i>	
calling_field	<i>field</i>	
called_field	<i>field</i>	
frequency_weight_field	<i>field</i>	
read_field_names	<i>boolean</i>	
coverage_threshold	<i>double</i>	Default value = 10. Min = 1 Max = 99.
min_group_size	<i>integer</i>	Default value = 2. Min = 2.
max_group_size	<i>integer</i>	Default value = 100. Min = 2.
calculate_statistics	<i>boolean</i>	

Chapter 2. Group analysis

Group analysis overview

Group analysis uses the interaction patterns of individuals in a network to identify groups of similar individuals. Characteristics of these groups influence the behavior of the individual group members. For example, small groups having many inter-member relationships and strong leaders have an increased risk of churn, even if no member of the group has actually churned.³ Predictive models that incorporate both group and individual measures will perform better than models that include only the latter.

Group analysis consists of the following general steps:

1. Determine relationship strengths that reflect social proximity. See the topic “Determining social similarity” for more information.
2. Partition the network into groups based on the relationship strength while obeying size restrictions. See the topic “Partitioning into groups” on page 10 for more information.
3. Profile both the groups and individuals, including identifying group leaders. See the topic “Describing groups and group members” on page 10 for more information.

Determining social similarity

Members of a group should be more similar to each other than they are to individuals who are not in the group. In network analysis, the similarity of two nodes depends on their relationships. For any node in a network, there is a set of nodes that are the targets of directed relationships with the node. For telecommunications data, this set corresponds to all of the people contacted by a particular individual. If two individuals contact the same set of people, those individuals are considered similar to each other. The more the sets of relationship targets overlap for two individuals, the more similar they are.

Consider the network shown in the "Example ten node network" figure.

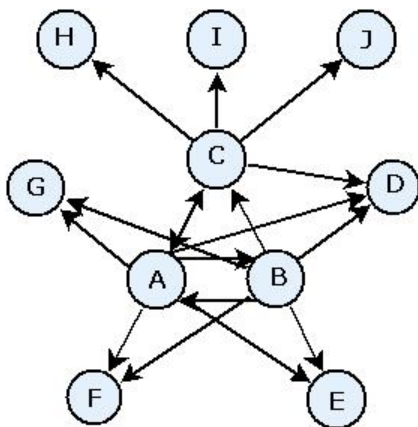


Figure 2. Example ten node network

The "Target nodes" table shows the target nodes for relationships originating at nodes A, B, and C.

3. Richter, Y., E. Yom-Tov, and N. Slonim. 2010. Predicting customer churn in mobile networks through analysis of social groups. In: *Proceedings of the 2010 SIAM international conference on data mining*. Columbus, OH: SDM 2010.

Table 7. Target nodes.

Source node	Target nodes
A	B, C, D, E, F, G
B	A, C, D, E, F, G
C	A, D, H, I, J

Nodes A and B have five common target nodes. Nodes A and C, on the other hand, have only one. Thus nodes A and B are more socially similar than nodes A and C.

Relationship weights such as call duration or frequency do not capture the similarity of nodes in a network. To include the notion of similarity, group analysis uses mutual information⁴ as the relationship weight. This statistic reflects the likelihood that two nodes are connected to the same node. If the relationships in the network have predefined weights, such as call frequency for telecommunications data, the mutual information incorporates those weights accordingly.

Partitioning into groups

Groups should consist of individuals that have high relationship weights with other group members, where the weights measure the similarity of the nodes in the relationship. As a result, group identification begins by omitting the weaker relationships in a network. The *coverage threshold* controls this process by defining the fraction of the strongest relationships to retain. For example, a coverage threshold value of 0.4 results in the strongest 40% of relationships being used for group identification, with the remaining 60% of the relationships omitted.

The remaining relationships may yield some very small or very large groups, which have limited predictive utility. To prevent such groups from being included in the analysis, you can specify minimum and maximum group sizes. Groups having sizes smaller than the minimum are omitted completely. Groups with sizes larger than the maximum, however, are divided into smaller groups within the acceptable size range. The groups remaining after dropping weak relationships and enforcing size limits are referred to as *core groups*.

Removing relationships from the original network may result in some individuals not being in any core group. However, those individuals may have connections with members of a group that warrant being included in the group. An individual is added to a group if there are many relatively strong relationships to the core members of the group, provided the group size limits are not violated. Thus, the final groups consist of a set of core members plus members added due to connectivity with the core.

Describing groups and group members

In addition to the density, indegree, and outdegree, other statistics describe group dynamics. In particular, authority and dissemination scores offer measures of the social status of the individuals within the groups. The role of each individual in a group is vitally important when trying to predict the behavior of both the group and its members.

The *authority score* for a node measures the tendency of other nodes in the group to connect to it. If many individuals are contacting a particular person, potentially asking for information or opinions, that person has the role of an authority. The authority scores for the nodes in group, which correspond to the stationary probabilities for a random walk with restarts through the group network, range from 0 to 1. The closer the authority score is to 1, the more authority that node has within the group. The node in the group with the highest authority score is referred to as the *authority leader* for the group. Dividing the largest score within a group by the smallest provides a measure of the overall strength of the authority leader.

4. Cover, T. M., and J. A. Thomas. 2006. *Elements of Information Theory, 2nd edition*. New York: John Wiley and Sons, Inc.

In contrast, the *dissemination score* for a node measures the tendency of the node to connect to other nodes in the group. If a particular person contacts many other people in the group, that person can significantly affect the opinions of the entire group. The dissemination scores for the nodes in group, which correspond to the stationary probabilities for a random walk with restarts through the group network in reverse, range from 0 to 1. The closer the dissemination score is to 1, the more the node connects to the other group members. The node in the group with the highest dissemination score is referred to as the *dissemination leader* for the group. Dividing the largest score within a group by the smallest provides a measure of the overall strength of the dissemination leader.

Group Analysis node

The Group Analysis node, which is available from the Sources palette, identifies sets of individuals in a network that are socially similar to each other and determines the relative social status of the individuals in the groups. Groups that are relatively small, have many connections between members, and include a strong leader are at an increased risk of churn due to group dynamics. You can use the results to determine group churn scores that, when combined with individual churn scores, improve the ability to predict churn over models based on individual scores alone.

Given the amount of data that typically comprises the network and the nature of the social interactions underlying the group definitions, group analysis is a time-consuming process. Typically, you would save the group analysis results to a database or file to be used as input to a predictive model. The group results would be refreshed relatively infrequently, such as monthly, while the model may be refreshed much more often.

Requirements. The node requires a fixed width text file defining the social network using three fields. One field identifies the source for each directed relationship, one field defines the destination for each directed relationship, and the third field specifies an optional strength for each relationship. All relationships in the network must be directional.

Specifying data for group analysis

The Data tab of the Group Analysis source node window enables you to specify the input file containing the network node relationships.

File. Specify the name of a file or folder containing the call detail records. You can enter a name or click the ellipsis button (...) to select a name from the file system. The path is shown once you have selected a name, and its contents are displayed with delimiters. If you specify a folder, the call detail records in all files contained in the folder are concatenated for the analysis; all files in the folder should have the same structure.

Read field names from file. Selected by default, this option treats the first row in the data file as names for the columns. If your first row is not a header, deselect this option to automatically give each field a generic name, such as *Field1* and *Field2*.

Network Definition Settings

The network definition settings define the roles for the fields.

Fields. Use the arrow buttons to manually assign items from this list to the various role fields. The icons indicate the valid measurement levels for each role field. Click the **All** button to select all the fields in the list, or click an individual measurement level button to select all fields with that measurement level.

Source. Select one field as the origin of the directed relationship.

Destination. Select one field as the target of the directed relationship.

Weight. Optionally, select one field corresponding to the number of times the relationship occurs in the network. The analysis weights the records accordingly when determining the relationship strengths.

Setting build options for group analysis

The Build Options tab of the Group Analysis source node window enables you to define options for identifying groups in the network.

Group Analysis Settings

The group analysis settings influence the size and relative strength of the groups formed.

Coverage Threshold. Define the fraction of the strongest network relationships to use in the analysis. For example, a coverage threshold of 0.2 indicates that only the relationships whose weights are among the top 20% of all weights will be used. This parameter, which ranges from 0 to 1, allows the analysis to focus on the strongest relationships, which should occur within groups. Higher threshold values yield groups with weaker relationships among their members. See the topic “Partitioning into groups” on page 10 for more information.

Min Group Size. Specify a lower boundary for the group size. Groups smaller than this value will not be returned.

Max Group Size. Specify an upper boundary for the group size. Groups larger than this value will be divided into smaller groups.

Calculate and Display Summary Statistics. If selected, the node computes and displays summary statistics for the analysis in addition to deriving the key performance indicator output. The calculation of these statistics can negatively impact the performance of the node for extremely large networks.

Viewing group analysis statistics

The Analysis tab of the Group Analysis source node window provides a summary overview of the groups identified in the network. The "Group Analysis summary statistics" table lists the available summary statistics for the groups.

Table 8. Group Analysis summary statistics.

Statistic	Description
Total Nodes in Groups	Number of nodes included in the identified groups
Total Links in Groups	Number of links included in the identified groups
Total Number of Groups	Number of groups identified in the network
Mean Group Size	Average number of nodes in a group
Mean Group Density	Average fraction of direct connections between nodes in a group. See the topic “Network density” on page 3 for more information.
Mean Fraction of Core Members	Average fraction of nodes in a group that are core nodes for the group. See the topic “Partitioning into groups” on page 10 for more information.
Mean Density of Core Group	Average fraction of direct connections between core nodes in a group
Mean InDegree	Average number of incoming links. See the topic “Nodal degree” on page 4 for more information.
Mean OutDegree	Average number of outgoing links. See the topic “Nodal degree” on page 4 for more information.

Select a specific statistic from the Summary Statistics table to view the distribution of values, the standard deviation, and the skewness for the statistic.

Reviewing these statistics assists in defining group analysis settings. Ideally, the results should show relatively small groups having high density values. For example, if there are some very large groups, consider reducing the maximum group size to divide them into smaller groups. Alternatively, if the group density values tend to be small, consider reducing the coverage threshold to focus on stronger relationships.

Output for group analysis

The Group Analysis node generates a variety of fields describing the groups and the individuals in the groups. You can augment existing models and data with these key performance indicators to improve the predictions generated by your models. For example, you can update individual churn propensity values to include group influences.

The "Key performance indicators for groups" table lists the key performance indicator fields for the groups identified in the analysis.

Table 9. Key performance indicators for groups.

Field	Description
GAG_GroupNumber	Unique identifier for a group
GAG_Size	Number of individuals in a group
GAG_Density	Fraction of direct connections between individuals in a group. See the topic "Network density" on page 3 for more information.
GAG_KernelDensity	Fraction of direct connections between core individuals in a group
GAG_CoreNodesFraction	Fraction of individuals in a group that are core individuals for the group. See the topic "Partitioning into groups" on page 10 for more information.
GAG_MaxRankType1	Maximum authority score of any group member. See the topic "Describing groups and group members" on page 10 for more information.
GAG_MinRankType1	Minimum authority score of any group member
GAG_MaxMinRankRatioType1	Ratio of the largest authority score to the smallest. This value reflects the authority strength of the group leader.
GAG_MaxRankType2	Maximum dissemination score of any group member. See the topic "Describing groups and group members" on page 10 for more information.
GAG_MinRankType2	Minimum dissemination score of any group member
GAG_MaxMinRankRatioType2	Ratio of the largest dissemination score to the smallest. This value reflects the dissemination strength of the group leader.

The "Key performance indicators for individuals" table lists the key performance indicator fields for the individuals in the network.

Table 10. Key performance indicators for individuals.

Field	Description
GAI_NodeNumber	Unique identifier for an individual
GAI_CoreNode	Indicator of whether the individual is a core individual for a group or not. See the topic "Partitioning into groups" on page 10 for more information.
GAI_RankType1	Authority score for the individual. See the topic "Describing groups and group members" on page 10 for more information.
GAI_RankOrderType1	Rank order in the group based on the authority scores

Table 10. Key performance indicators for individuals (continued).

Field	Description
GAI_RankType2	Dissemination score for the individual. See the topic “Describing groups and group members” on page 10 for more information.
GAI_RankOrderType2	Rank order in the group based on the dissemination scores
GAI_InDegree	Number of relationships in which the individual is the target of the relationship. See the topic “Nodal degree” on page 4 for more information.
GAI_OutDegree	Number of relationships in which the individual is the source of the relationship. See the topic “Nodal degree” on page 4 for more information.
GAI_GroupLeaderType1	Whether the node is an authority leader, whose leadership score is derived from incoming links. See the topic “Describing groups and group members” on page 10 for more information.
GAI_GroupLeaderConfidenceType1	The confidence that the node is an authority leader.
GAI_GroupLeaderType2	Whether the node is a dissemination leader, whose leadership score is derived from outgoing links. See the topic “Describing groups and group members” on page 10 for more information.
GAI_GroupLeaderConfidenceType2	The confidence that the node is a dissemination leader.

Chapter 3. Diffusion analysis

Diffusion analysis overview

Diffusion Analysis identifies individuals that are most affected by other individuals in a social network, quantifying the effect as *diffused energy*. The process uses a spreading activation approach, in which an effect iteratively spreads from network nodes to their immediate neighbors, diminishing in size as it moves from node to node⁵. Upon receiving energy, a node becomes activated and transmits a portion of that energy to any neighbors who are targets of directed relationships with the node.

The *spreading factor* defines the proportion of energy transmitted by an activated node, with the remaining amount retained by the node. Any nodes receiving this energy will themselves transmit the same proportion to their neighbors, resulting in a decaying process for energy transmission. Large spreading factors correspond to more energy being sent, allowing energy to reach nodes more distant from the initial activated nodes before the process decays completely. Small spreading factors result in diffusion processes that decay rapidly, with the transmitted energy remaining relatively close to the initial nodes.

The total amount of energy being diffused by an activated node is distributed among all nodes who are targets of directed relationships with the node. The amount each node receives depends on the strength of the relationship with the activated node. The fraction sent to a particular node equals the relationship weight divided by the total of the weights for all relationships in which the activated node is the source. Consequently, neighbors having relationships with higher relative weights receive more energy than neighbors with lower relative weights.

The diffusion process stops when one of the following conditions occurs:

- activated nodes are not the source of any directed relationships
- the amount of energy being transmitted is below the *accuracy threshold*, a limit on the amount of energy being transferred for the process to continue
- the number of iterations reaches a specified limit

When diffusion completes, the nodes having the most diffused energy are the most sensitive to the effect that initiated the process. For example, if the process begins at nodes that churn, the nodes with the highest energy are the most at risk of churning themselves. You can give those nodes special attention to prevent them from churning.

Diffusion analysis example

Consider the network shown in “Displaying networks” on page 2. The network consists of seven nodes having directed relationships of varying strengths with each other.

The “Example diffusion process” table illustrates a diffusion process across the network using a spreading factor of 0.80. Initially, node A contains all of the energy, arbitrarily assigned a value of 1.00. In step 1, this node activates, spreading 80% of its energy to the three neighbors that are targets of relationships while retaining 20% for itself. The relationships have a total weight of 100. The relationship with node D accounts for half of this total so this node receives half of the diffused energy, or 0.40. The relationship with node B accounts for 20% of the total weight, yielding in a diffused energy value of 0.16. Node C receives the remaining energy, 0.24, which is 30% of the amount originating from node A.

5. Dasgupta, K., R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjee, A. A. Nanavati, and A. Joshi. 2008. Social ties and their relevance to churn in mobile telecom networks. In: *Proceedings of the 11th international conference on extending database technology*. New York, NY: ADM.

Table 11. Example diffusion process.

Step	Node A	Node B	Node C	Node D	Node E	Node F	Node G
0	1.00	0	0	0	0	0	0
1	0.20	0.16	0.24	0.40	0	0	0
2	0.20	0.16	0.05	0.08	0.08	0.24	0.19
3	0.20	0.16	0.05	0.08	0.08	0.08	0.38

Having received energy in step 1, nodes B, C, and D activate in step 2 and diffuse energy to any target neighbors. Node B has no target nodes so it diffuses no energy. Node C, on the other hand, has a target and diffuses 80% of its energy to that node, G. Finally, Node D spreads 80% of its energy to nodes E and F, with F receiving three times the amount of E due to its stronger relationship with D.

In step 3, nodes E, F, and G activate. Nodes E and G have no target nodes so neither diffuses any energy. However, node F diffuses 80% of its energy to its single target node, G. This node now has a total energy value of 0.38 consisting of energy diffused from both nodes C and F.

At this point, node G activates but has no target nodes for diffusion so the process terminates. The energy originating at node A has spread across the network, with node G receiving the highest amount. If the energy introduced in the network represents churn, node G would be most affected by node A churning.

Diffusion Analysis node

The Diffusion Analysis node, which is available from the Sources palette, propagates an effect from a specified set of individuals across a social network, using the network relationships to identify the individuals most impacted by the effect. If the effect is churn, for instance, the node identifies those individuals most likely to churn due to other specific individuals in the network churning. You can augment existing models and data with the node output to improve the predictions generated by those models. For example, you can update individual churn propensity values to include diffusion influences.

Requirements. The node requires two fixed width text files containing the data to be analyzed. The first file defines the social network using three fields. One field identifies the source for each directed relationship, one field defines the destination for each directed relationship, and the third field specifies an optional strength for each relationship. All relationships in the network must be directional. The second file contains a list of identifiers from which the effect should begin.

Specifying data for diffusion analysis

The Data tab of the Diffusion Analysis source node window enables you to specify the input file containing the network node relationships.

File. Specify the name of a file or folder containing the call detail records. You can enter a name or click the ellipsis button (...) to select a name from the file system. The path is shown once you have selected a name, and its contents are displayed with delimiters. If you specify a folder, the call detail records in all files contained in the folder are concatenated for the analysis; all files in the folder should have the same structure.

Read field names from file. Selected by default, this option treats the first row in the data file as names for the columns. If your first row is not a header, deselect this option to automatically give each field a generic name, such as *Field1* and *Field2*.

Network Definition Settings

The network definition settings define the roles for the fields.

Fields. Use the arrow buttons to manually assign items from this list to the various role fields. The icons indicate the valid measurement levels for each role field. Click the **All** button to select all the fields in the list, or click an individual measurement level button to select all fields with that measurement level.

Source. Select one field as the origin of the directed relationship.

Destination. Select one field as the target of the directed relationship.

Weight. Select one field representing the relationship weight. For call detail records, the weight may correspond to any of a number of call characteristics reflecting the relationship strength, such as call duration or call frequency.

Setting build options for diffusion analysis

The Build Options tab of the Diffusion Analysis source node window enables you to define options for the diffusion process.

Diffusion Seed List

The diffusion seed list identifies the network nodes from which energy should begin diffusing. For example, in churn analysis, this list identifies the network nodes that have already churned. Alternatively, this list could contain nodes at the highest risk of churn to explore the effects on the network should those nodes actually churn.

File. Specify the name of the text file containing the node identifiers, with each identifier corresponding to a line in the file. You can enter a filename or click the ellipsis button (...) to select a file.

Diffusion Analysis Settings

The diffusion analysis settings determine how aggressive the diffusion is and when the process stops.

Spreading Factor. Define the fraction of energy a node transfers at each step of the diffusion process. Smaller values limit the energy to the nodes nearest to the node that is diffusing the energy. Conversely, larger values permit the energy to affect more distant nodes.

Maximum Iterations. Specify an upper boundary for the number of diffusion iterations. The diffusion process automatically terminates if this limit is reached.

Accuracy Threshold. Specify the smallest change in node energy that warrants continuing the diffusion process. If the change in energy for every node is smaller than this value, the diffusion process terminates.

Calculate and Display Summary Statistics. If selected, the node computes and displays summary statistics for the analysis in addition to deriving the key performance indicator output. The calculation of these statistics can negatively impact the performance of the node for extremely large networks.

Viewing diffusion analysis statistics

The Analysis tab of the Diffusion Analysis source node window provides a summary overview of the diffusion results. The "Diffusion Analysis summary statistics" table lists the available summary statistics.

Table 12. Diffusion Analysis summary statistics.

Statistic	Description
Total Nodes in Network	Number of nodes in the network
Total Links in Network	Number of links in the network
Total Diffusion Seeds in Network	Number of nodes used as seeds for the diffusion process

Table 12. Diffusion Analysis summary statistics (continued).

Statistic	Description
Mean Influence	Average amount of diffused energy associated with individuals.
Mean InDegree	Average number of relationships in which an individual is the target of the relationship. See the topic "Nodal degree" on page 4 for more information.
Mean OutDegree	Average number of relationships in which an individual is the source of the relationship. See the topic "Nodal degree" on page 4 for more information.

Select a specific statistic from the Summary Statistics table to view the distribution of values, the standard deviation, and the skewness for the statistic.

Output for diffusion analysis

The Diffusion Analysis node generates a variety of fields describing the individuals in the network. You can augment existing models and data with these key performance indicators to improve the predictions generated by your models. For example, you can update individual churn propensity values to include diffusion influences.

The "Diffusion Analysis key performance indicators" table lists the key performance indicator fields for the individuals in the analysis.

Table 13. Diffusion Analysis key performance indicators.

Field	Description
DA_NodeNumber	Unique identifier for an individual
DA_DiffusedEnergy	Amount of diffused energy associated with the individual. For churn analysis, higher values indicate a greater propensity to churn than lower values.
DA_InDegree	Number of relationships in which the individual is the target of the relationship. See the topic "Nodal degree" on page 4 for more information.
DA_OutDegree	Number of relationships in which the individual is the source of the relationship. See the topic "Nodal degree" on page 4 for more information.

Notices

This information was developed for products and services offered worldwide.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who want to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, and [ibm.com](http://www.ibm.com) are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Other product and service names might be trademarks of IBM or other companies.

Glossary

A

accuracy threshold . A stopping criterion for a diffusion process corresponding to the smallest change in node energy that warrants continuing diffusion.

authority leader . The individual in a group having the highest authority score.

authority score . Measure indicating how much the other members of a group connect to an individual. Authority scores associate the importance of an individual with the number of relationships ending at him or her.

C

core group . A group of nodes remaining after omitting weaker relationships from a network and applying group size limits.

coverage threshold . The fraction of the strongest relationships used for group identification.

D

degree . Measure of node activity in a network, defined as the total number of relationships involving the node.

density . A measure of network cohesion defined as the number of observed relationships divided by the number of possible relationships.

dichotomous relationship . A connection between two nodes that can take only one of two values, indicating presence or absence of the connection.

diffused energy . Effect originating from specific nodes in a network that spreads across the entire network, decreasing in size as the distance from the initial nodes gets larger.

directional relationship . A connection between two nodes that originates from one (the source) and ends at the other (the destination).

dissemination leader . The individual in a group having the highest dissemination score.

dissemination score . Measure indicating how much an individual connects to other members of a group. Dissemination scores associate the importance of an individual with the number of relationships originating from him or her.

I

indegree . Measure of prestige for a node in a network consisting of directed relationships, defined as the total number of relationships in which the node is the target.

N

nondirectional relationship . An association between two nodes.

O

outdegree . Measure of centrality for a node in a network consisting of directed relationships, defined as the total number of relationships in which the node is the source.

S

sociogram . A visual representation of a social network in which lines connecting points depict individuals having a relationship with each other.

spreading factor . The fraction of energy a node transfers to its neighbors in a diffusion process step.

V

valued relationship . An connection between two nodes that has an associated weight indicating the strength of the relationship.

Index

A

accuracy threshold 15, 17
analysis tab 5
 diffusion analysis 17
 group analysis 12
analyzing data 6
annotations tab 5
authority leaders 10, 13
authority scores 10, 13

B

build options tab 5
 diffusion analysis 17
 group analysis 12

C

call detail records 7
clearing results 6
core groups 10
core nodes 12, 13
coverage threshold 10, 12

D

data tab 5
 diffusion analysis 16
 group analysis 11
degree 4
density 3, 12, 13
destination fields 7
dichotomous relationships 1
diffused energy 17, 18
diffusion analysis 5, 15
 accuracy threshold 15, 17
 options 16
 requirements 16
 seeds 17
 spreading factor 15, 17
diffusion seed list 17
directional relationships 1
dissemination leaders 10, 13
dissemination scores 10, 13

E

executing streams
 using IBM SPSS Modeler Server Social
 Network Analysis 5

F

filter tab 5
filtering output 5

G

GAG_GroupNumber 13
group analysis 5
 coverage threshold 10, 12
 group sizes 10, 12
 requirements 11
group sizes 12, 13
 limits 10, 12

I

indegree 4, 12, 13, 17, 18

K

kernel density 12, 13
key performance indicators 13, 18

M

mean core group density 12
mean core member fraction 12
mean group density 12
mean group size 12
mean indegree 12, 17
mean influence 17
mean outdegree 12, 17
measurement levels 5

N

nondirectional relationships 1

O

outdegree 4, 12, 13, 17, 18

P

previewing output 5
properties
 scripting 7

R

relationships
 direction 1
 valued 1

S

scripting
 properties 7
seeds
 for diffusion analysis 17
source fields 7
spreading factor 15, 17

T

types tab 5

V

valued relationships 1

W

weight fields 7, 11, 16



Printed in USA