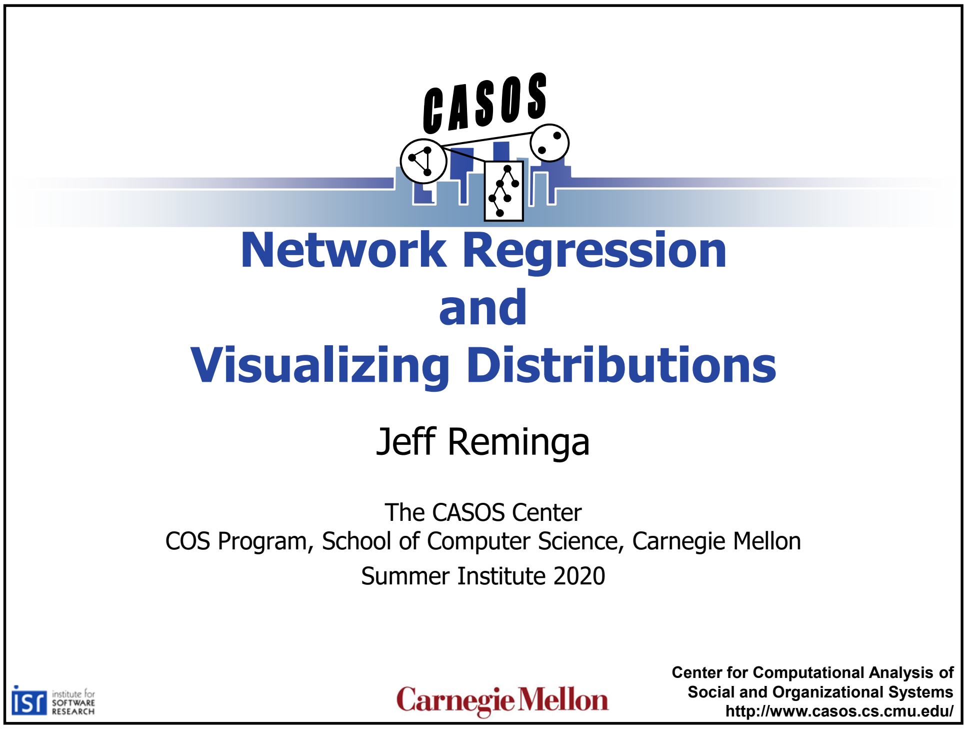
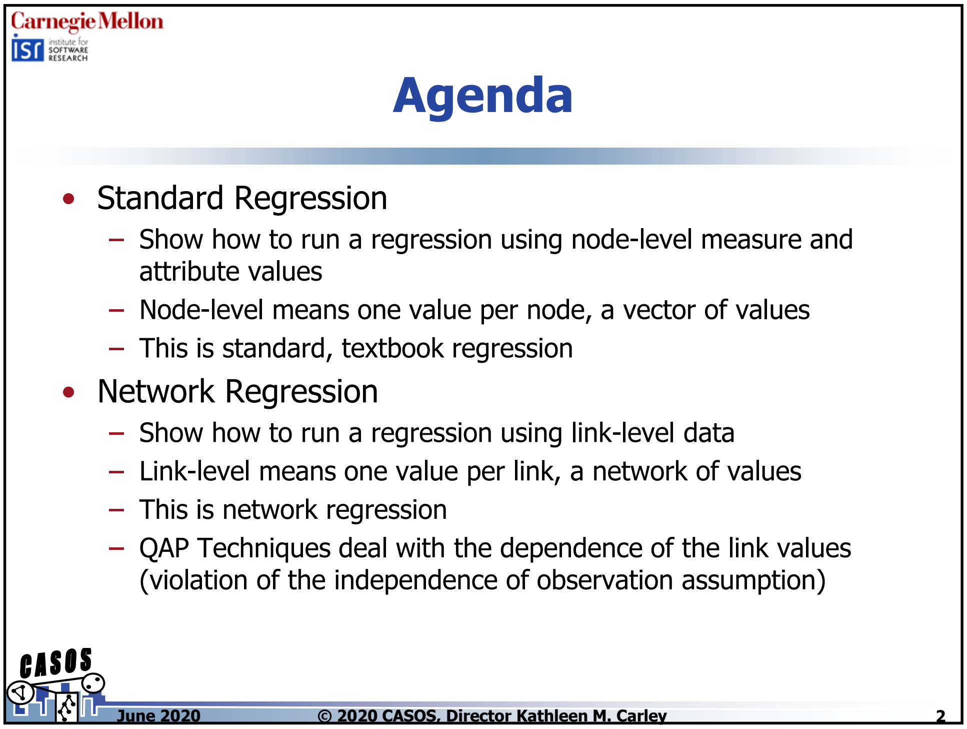


Network Regression and Visualizing Distributions

Jeff Reminga

The CASOS Center
COS Program, School of Computer Science, Carnegie Mellon
Summer Institute 2020

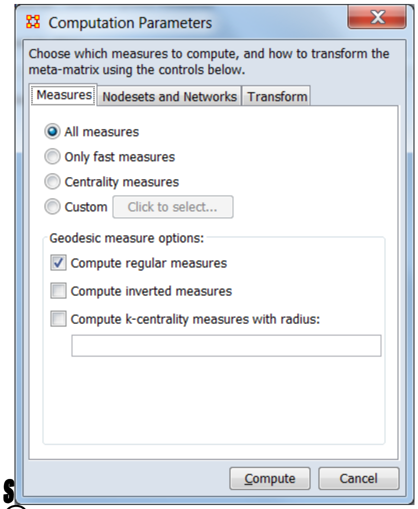
Center for Computational Analysis of
Social and Organizational Systems
http://www.casos.cs.cmu.edu/



Agenda

- Standard Regression
  - Show how to run a regression using node-level measure and attribute values
  - Node-level means one value per node, a vector of values
  - This is standard, textbook regression
- Network Regression
  - Show how to run a regression using link-level data
  - Link-level means one value per link, a network of values
  - This is network regression
  - QAP Techniques deal with the dependence of the link values (violation of the independence of observation assumption)

Standard Regression



Choose which measures to compute, and how to transform the meta-matrix using the controls below.

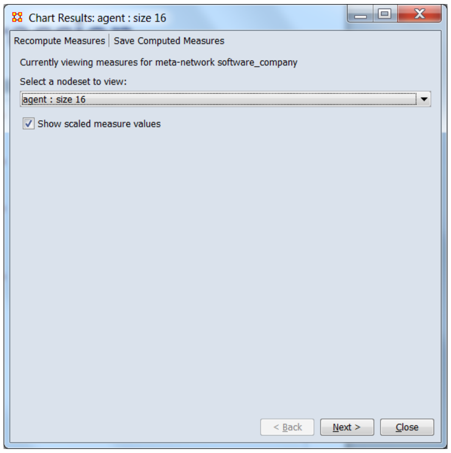
● All measures
○ Only fast measures
○ Centrality measures
○ Custom

Geodesic measure options:

☑ Compute regular measures
□ Compute inverted measures
□ Compute k-centrality measures with radius:

- Load the "softco.xml" dataset
- Select the softco meta-network in the Meta-Network Manager
- Click the main menu item Visualizations\Measure Charts (or Measure Charts button)
- Select the measures to compute: All, Fast, Centrality, or custom
- Geodesic measures (betweenness, closeness, path lengths) have options:
  - Use Inverted measures when the link values are dissimilarities (such as distance)
  - Use k-centrality for large data: as only considering neighbors within distance k
  - Inverted must be used needed when the links mean "matter of interpretation of link

Standard Regression…



Currently viewing measures for meta-network software_company

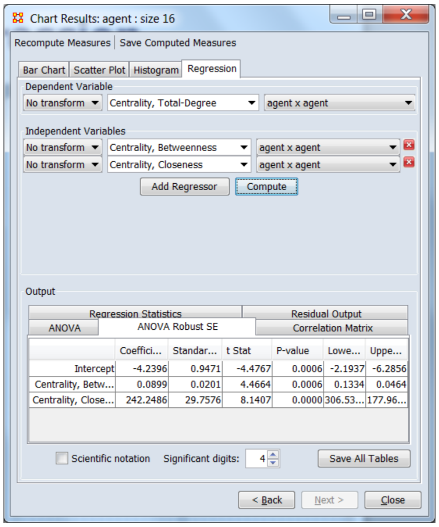
Select a nodeset to view:

agent : size 16

☑ Show scaled measure values

- After computing measures, you will see this dialog where you select which nodeset values to view
- Click whether to use scaled or unscaled values
- Click next

Standard Regression…



Dependent Variable: No transform | Centrality, Total-Degree | agent x agent

Independent Variables:
No transform | Centrality, Betweenness | agent x agent
No transform | Centrality, Closeness | agent x agent

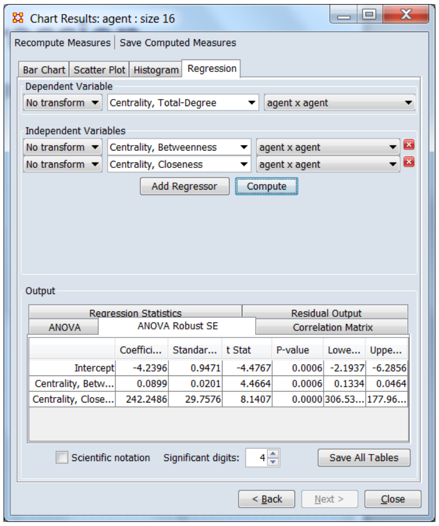
Output

	Coeffici...	Standar...	t Stat	P-value	Lowe...	Uppe...
Intercept	-4.2396	0.9471	-4.4767	0.0006	-2.1937	-6.2856
Centrality, Betw...	0.0899	0.0201	4.4664	0.0006	0.1334	0.0464
Centrality, Close...	242.2486	29.7576	8.1407	0.0000	306.53...	177.96...

☐ Scientific notation Significant digits: 4

- Choose the Regression tab at the top
- The independent variable and the dependent variables are all node-level measures or numeric attributes
- In this regression, the dependent variable is the measure Total Degree Centrality computed on the Agent x Agent network
- The dependent variables are Betweenness and Closeness Centrality also computed on the Agent x Agent network.
- Regression results are print below
- Use the Save All Tables to save values into a CSV file

Standard Regression…



Dependent Variable: No transform | Centrality, Total-Degree | agent x agent

Independent Variables:
No transform | Centrality, Betweenness | agent x agent
No transform | Centrality, Closeness | agent x agent

Output

	Coeffici...	Standar...	t Stat	P-value	Lowe...	Uppe...
Intercept	-4.2396	0.9471	-4.4767	0.0006	-2.1937	-6.2856
Centrality, Betw...	0.0899	0.0201	4.4664	0.0006	0.1334	0.0464
Centrality, Close...	242.2486	29.7576	8.1407	0.0000	306.53...	177.96...

☐ Scientific notation Significant digits: 4

- Regression results are print below
- The results indicate that Total Degree Centrality is determined more by Closeness Centrality than Betweenness Centrality
- This makes sense because an increase in closeness would entail an increase in degree.
- Use the Save All Tables to save values into a CSV file

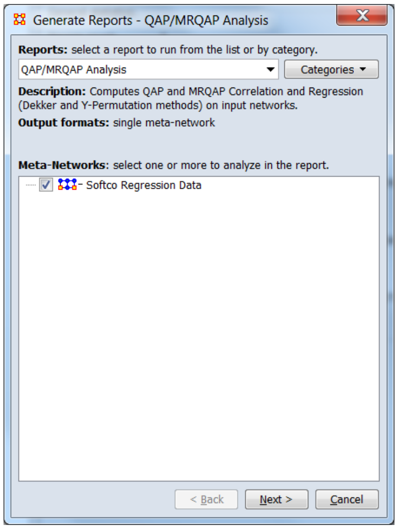
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Network Regression

- In Standard Regression the unit of analysis is a node, and node-level measure and numeric attribute data was used
- In Network Regression the unit of analysis is the dyad (a pair of nodes), and ORA lets you input three types of data:
  - A network directly, such as an Agent x Agent network
  - A vector of node-level numeric attributes (such the Age of each node) repeated by row or column to form a network
  - A vector of node-level measure values (such as Betweeenness Centrality) repeated by row or column to form a network
- Mathematically, the same calculations are used as those in node-level regression
- The link values are "stretched out" into a vector and then input into the standard regression routines
- Special techniques are used to handle the lack of independence of observations in the link-level observations, namely, Multiple Regression Quadratic Assignment Procedure (MRQAP)

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Network Regression…



Reports: select a report to run from the list or by category.

QAP/MRQAP Analysis

Description: Computes QAP and MRQAP Correlation and Regression (Dekker and Y-Permutation methods) on input networks.

Output formats: single meta-network

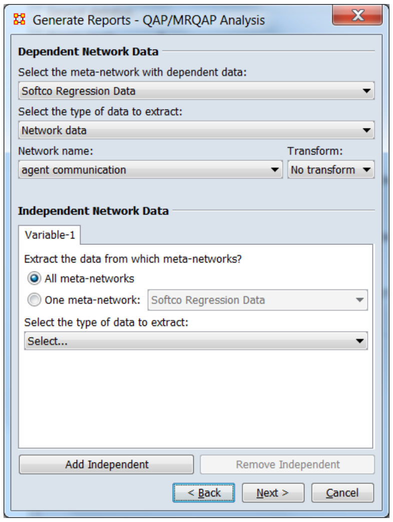
Meta-Networks: select one or more to analyze in the report.

☑ Softco Regression Data

- As an example, load the "softco-regression.xml" dataset
- Click Generate Reports
- Select the QAP/MRQAP Analysis report

Carnegie Mellon

Network Regression…


Generate Reports - QAP/MRQAP Analysis

Dependent Network Data

Select the meta-network with dependent data: Softco Regression Data

Select the type of data to extract: Network data

Network name: agent communication

Transform: No transform

Independent Network Data

Variable-1

Extract the data from which meta-networks?

● All meta-networks

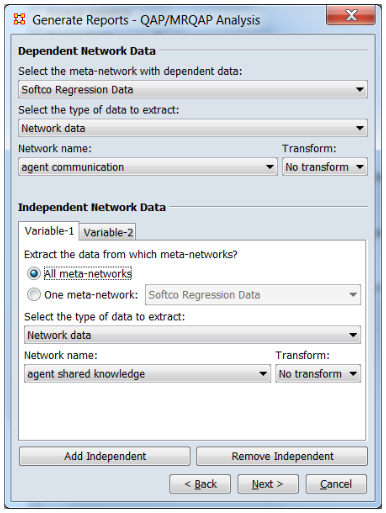
○ One meta-network: Softco Regression Data

Select the type of data to extract: Select...

- Click next several times
- We will assess to what degree agent communication is determined by agent shared knowledge and agent shared tasks
- First select the Dependent Data as the Network data "agent communication"
- We could transform each link value using the log, sqrt, etc. if we thought that made sense

Carnegie Mellon

Network Regression…


Generate Reports - QAP/MRQAP Analysis

Dependent Network Data

Select the meta-network with dependent data: Softco Regression Data

Select the type of data to extract: Network data

Network name: agent communication

Transform: No transform

Independent Network Data

Variable-1 Variable-2

Extract the data from which meta-networks?

● All meta-networks

○ One meta-network: Softco Regression Data

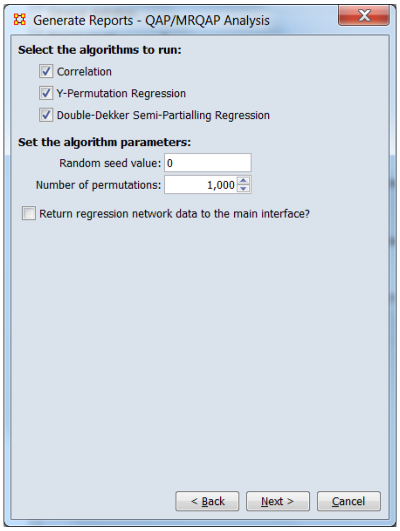
Select the type of data to extract: Network data

Network name: agent shared knowledge

Transform: No transform

- Now select the Independent Data
- In the Variable-1 tab, select "Network data" type and the "agent shared knowledge" network
- Then click the "Add Independent" button
- In the Variable-2 tab, select "Network data" type and the "agent shared task" network
- Click Next

Network Regression…



Generate Reports - QAP/MRQAP Analysis

Select the algorithms to run:

- ☑ Correlation
- ☑ Y-Permutation Regression
- ☑ Double-Dekker Semi-Partialling Regression

Set the algorithm parameters:

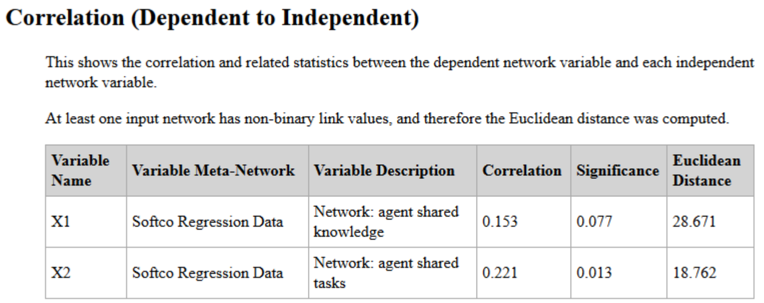
Random seed value: 0

Number of permutations: 1,000

□ Return regression network data to the main interface?

- After clicking Next, we have some options to select
- Correlation will compute the correlation between the dependent and independent variables
- Y-Permutation Regression will compute regression using a QAP technique involving permuting the independent network
- Double-Dekker Semi-Partialling Regression will compute the regression using another technique to deal with the inherent dependence of link data
- Correlation and Regression will use a bootstrapping technique that inputs a random seed and number of iterations

Network Regression…



Correlation (Dependent to Independent)

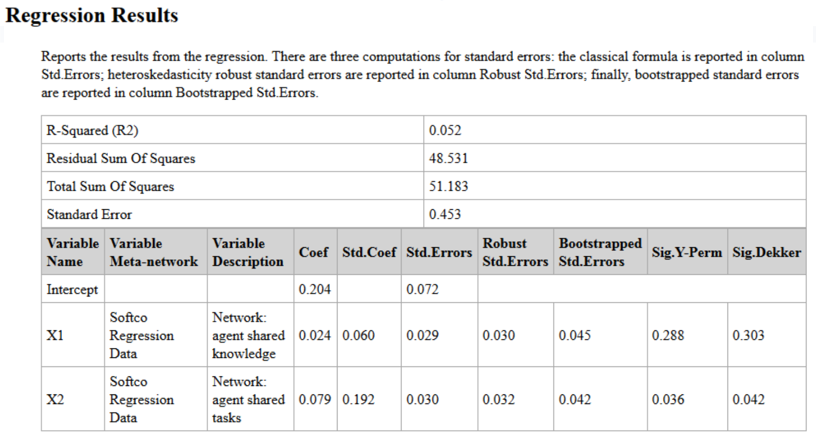
This shows the correlation and related statistics between the dependent network variable and each independent network variable.

At least one input network has non-binary link values, and therefore the Euclidean distance was computed.

Variable Name	Variable Meta-Network	Variable Description	Correlation	Significance	Euclidean Distance
X1	Softco Regression Data	Network: agent shared knowledge	0.153	0.077	28.671
X2	Softco Regression Data	Network: agent shared tasks	0.221	0.013	18.762

- Click Next, choose an output file, and click Finish
- This is the Correlation output from the report
- Note that Agent Shared Tasks is more highly correlated than shared Knowledge, but with less significance
- Euclidean distance is used because the independent networks are both valued

Network Regression…



Regression Results

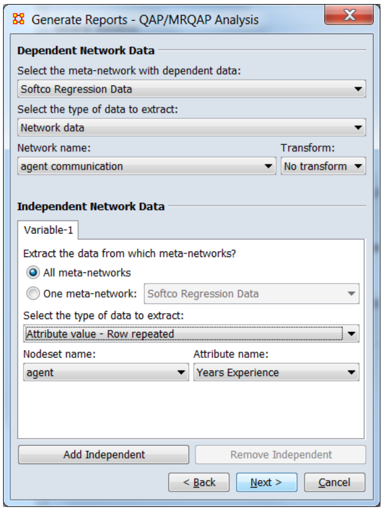
Reports the results from the regression. There are three computations for standard errors: the classical formula is reported in column Std.Errors; heteroskedasticity robust standard errors are reported in column Robust Std.Errors; finally, bootstrapped standard errors are reported in column Bootstrapped Std.Errors.

R-Squared (R2)	0.052
Residual Sum Of Squares	48.531
Total Sum Of Squares	51.183
Standard Error	0.453

Variable Name	Variable Meta-network	Variable Description	Coef	Std.Coef	Std.Errors	Robust Std.Errors	Bootstrapped Std.Errors	Sig.Y-Perm	Sig.Dekker
Intercept			0.204		0.072				
X1	Softco Regression Data	Network: agent shared knowledge	0.024	0.060	0.029	0.030	0.045	0.288	0.303
X2	Softco Regression Data	Network: agent shared tasks	0.079	0.192	0.030	0.032	0.042	0.036	0.042

- The Regression results show as well that Shared Tasks has a greater effect on communication than Shared Knowledge
- Again, however, the significance of Shared Tasks using both using Y-Permutation and Double Dekker Semi-Partialling is less

Network Regression: Attributes



Generate Reports - QAP/MRQAP Analysis

Dependent Network Data

Select the meta-network with dependent data: Softco Regression Data

Select the type of data to extract: Network data

Network name: agent communication

Transform: No transform

Independent Network Data

Variable-1

Extract the data from which meta-networks?

● All meta-networks

○ One meta-network: Softco Regression Data

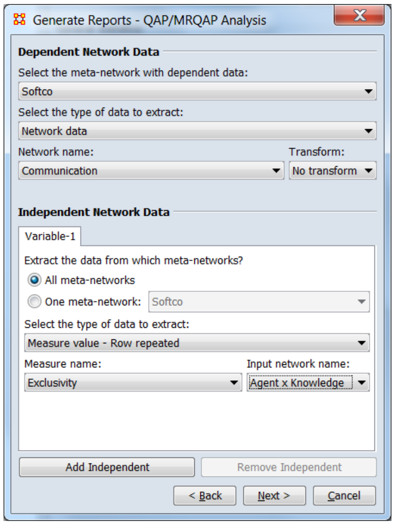
Select the type of data to extract: Attribute value - Row repeated

Nodeset name: agent

Attribute name: Years Experience

- Another example is to assess the degree to which a communication between two agents is explained by the years experience of the actors.
- Use *Attribute value - Row Repeated* to compare the link value (a,b) with the attribute value of node b
- Use *Attribute value – Column Repeated* to compare the link value (a,b) with the attribute value of a
- Select the Independent variable as shown to use the Years Experience attribute

Network Regression: Measures



Generate Reports - QAP/MRQAP Analysis

Dependent Network Data

Select the meta-network with dependent data: Softco

Select the type of data to extract: Network data

Network name: Communication

Transform: No transform

Independent Network Data

Variable-1

Extract the data from which meta-networks?

● All meta-networks

○ One meta-network: Softco

Select the type of data to extract: Measure value - Row repeated

Measure name: Exclusivity

Input network name: Agent x Knowledge

- Another example, using the "softco.xml" dataset, is to assess the degree to which a communication from agent A → B is explained by exclusive access of B to knowledge.
- That is, to what extent there is a communication from A → B so that A can access the knowledge of B.
- Use *Measure value - Row Repeated* to compare the link value (a,b) with the attribute value of node b
- Select the Independent variable as shown to use the Exclusivity measure computed on Agent x Knowledge

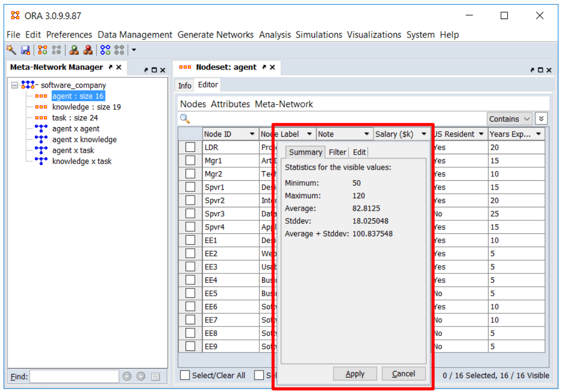
Visualizing Distributions

- Nodesets can have node attributes
- The distribution of node attribute values can be shown for categorical number, categorical text, or numerical data types
- Networks have link values whose distribution can be shown

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Node Attribute Distribution – Number values

Using the software company dataset, display the Agent nodeset in the editor, go to the Edit tab, and the click on down arrow for the Salary ($k) column



Statistics for the visible values:

Minimum:	50
Maximum:	120
Average:	82.8125
Stddev:	18.025048
Average + Stddev:	100.837548

	Node ID	Node Label	US Resident	Years Exp...
□	LDR	Pro	Yes	20
□	Mgr1	Art	Yes	15
□	Mgr2	Tec	Yes	10
□	Spvr1	Des	Yes	15
□	Spvr2	Int	Yes	20
□	Spvr3	Dat	No	25
□	Spvr4	App	Yes	15
□	EE1	Des	Yes	10
□	EE2	We	Yes	5
□	EE3	Usa	Yes	5
□	EE4	Bus	Yes	5
□	EE5	Bus	No	5
□	EE6	Sof	Yes	10
□	EE7	Sof	No	10
□	EE8	Sof	No	5
□	EE9	Sof	No	5

0 / 16 Selected, 16 / 16 Visible

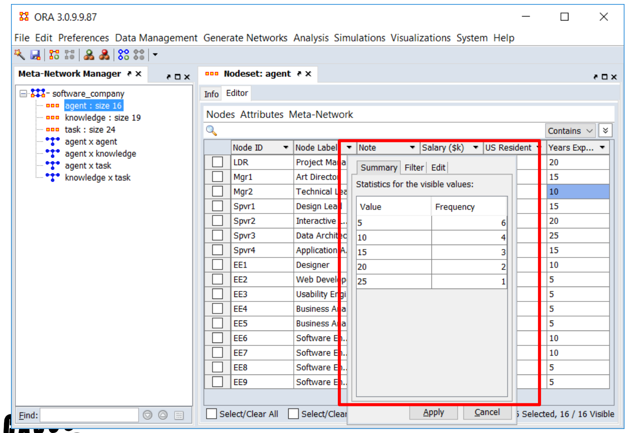
In red is the distribution of values shown when the array is clicked.

The min, max, mean, stddev of the salaries is shown because this is a numerical attribute

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Node Attribute Distribution – Categorical Number values

Now click the down arrow for the Years Experience column



Statistics for the visible values:

Value	Frequency
5	6
10	4
15	3
20	2
25	1

	Node ID	Node Label	Years Exp...
□	LDR	Project Ma	20
□	Mgr1	Art Directo	15
□	Mgr2	Technical Le	10
□	Spvr1	Design Lea	15
□	Spvr2	Interactive ...	20
□	Spvr3	Data Architec	25
□	Spvr4	Application A.	15
□	EE1	Designer	10
□	EE2	Web Develop	5
□	EE3	Usability Engi	5
□	EE4	Business A	5
□	EE5	Business A	5
□	EE6	Software E...	10
□	EE7	Software E...	10
□	EE8	Software E...	5
□	EE9	Software E...	5

5 Selected, 16 / 16 Visible

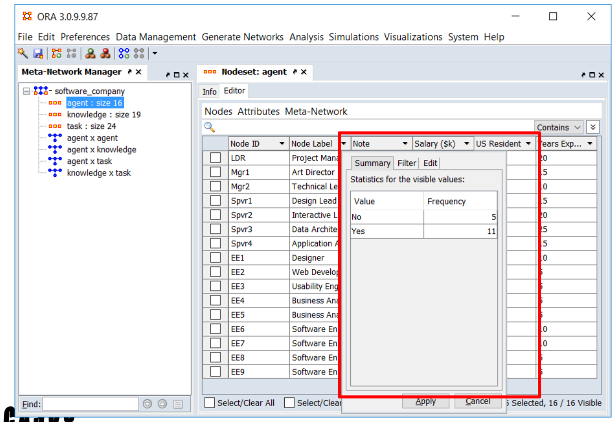
The values are shown as a frequency table because Years Experience is number category variable.

The data type of an attribute can be changed using the **Edit tab** (Summary | Filter | Edit)

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Node Attribute Distribution – Categorical Text values

Now click the down arrow for the US Resident column



Node ID	Node Label	Note	Salary ($k)	US Resident	Years Exp...
LDR	Project Mana				
Mgr1	Art Director				
Mgr2	Technical Le				
Spvr1	Design Lead				
Spvr2	Interactive L				
Spvr3	Data Archite				
Spvr4	Application A				
EE1	Designer				
EE2	Web Develo				
EE3	Usability Eng				
EE4	Business Ana				
EE5	Business Ana				
EE6	Software En				
EE7	Software En				
EE8	Software En				
EE9	Software En				

Statistics for the visible values:

Value	Frequency
No	5
Yes	11

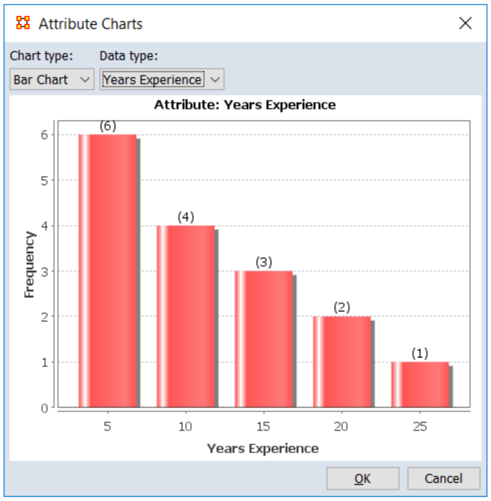
The values are shown as a frequency table because US Resident is text category variable.

The data type of an attribute can be changed using the **Edit tab**

(Summary | Filter | Edit)

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Node Attribute Distribution - Charts



Attribute Charts

Chart type: Bar Chart

Data type: Years Experience

- The Nodeset Editor also contains a charting feature to show the distribution of values
- Go to the nodeset editor's main menu: Attributes and select "Charts" (at bottom)
- Shown is the bar chart for Years Experience

Visualizing Network Value Distributions

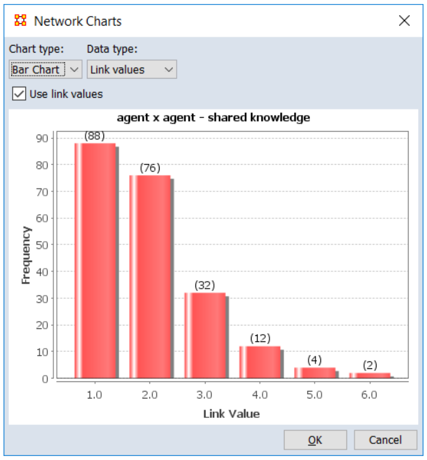


Chart type: Bar Chart

Data type: Link values

☑ Use link values

- The network editor has a similar chart tool to display the distribution of row sum values, column sum values, and links values
- This is located under the Network Editor's main menu: Nodes \ Charts
- Shown is the distribution of link values for the shared knowledge (folded agent x knowledge) network