



Handling Large Datasets

Jeff Reminga

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



Carnegie Mellon

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



Overview

- Importing files that have millions of links
- Creating subsets of data using:
 - Components
 - K-cores
- Selecting which measures to run in reports
 - Identify memory and time intensive measures in ORA
 - Do not compute them in reports and charts
- Selecting networks to analyze in reports and charts
- Selecting sphere of influence and paths to visualize



June 2020

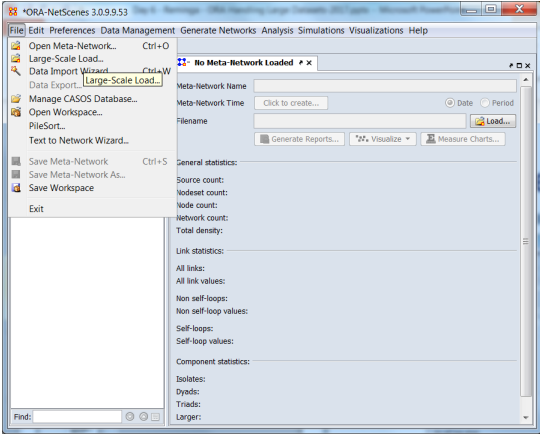
© 2020 CASOS, Director Kathleen M. Carley

2



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Large Scale Loader

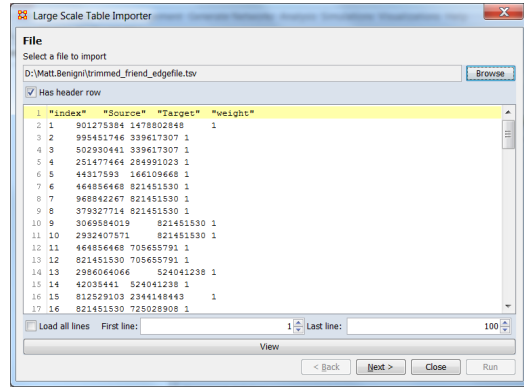


- Imports delimited files (e.g. comma or tab separated files) that contain hundreds of millions of lines
- File \ Large-Scale Load

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 3

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Large Scale Loader: File



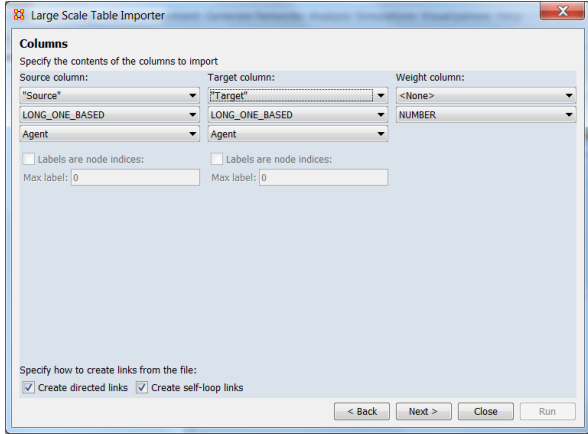
- Go to the Day 6 folder called Large Scale
- Browse to the file: trimmed_friend_edgfile.tsv
- The first 100 lines of the file are previewed
- Click on **Has header row**
- Use the controls at the bottom to view other portions of the file

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 4



Carnegie Mellon
IST Institute for Software Research

Large Scale Loader: Columns



This tool is for loading a single network

Each line of the file will create a link from the node in the source column to the node in the target column with the specified weight

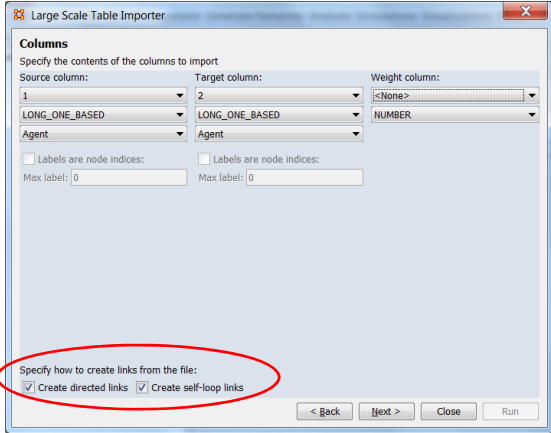
Large scale files will often integers or longs (big integers)

Select the type of data in each column

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 5

Carnegie Mellon
IST Institute for Software Research

Large Scale Loader: Columns



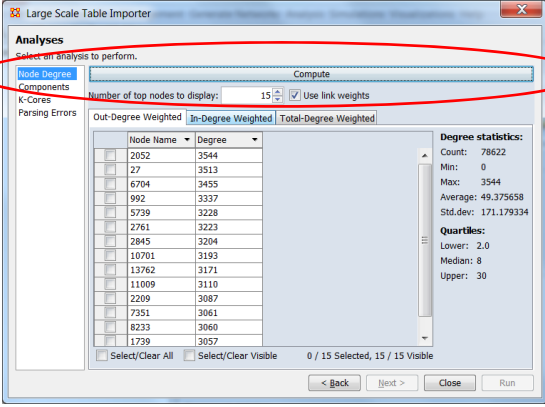
- The weight column is optional – it is more space efficient to not have weights
- Choose at the bottom whether links are to be considered directed and whether self-loops should be allowed
- Click Next

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 6



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Large Scale Loader: Node Degree Analyses

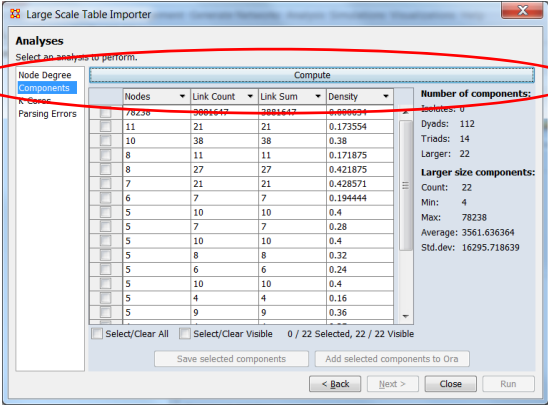


- The main idea when handling large scale data is to pre-select what you want to load
- Analyses let you identify nodes by degree, component, or k-core and load them
- Select Node Degree and the Compute bar
- The nodes with highest degree are shown and statistics are shown on the right hand side

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 7

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Large Scale Loader: Component Analyses



- Network components can also be investigated
- Click on Components and then the Compute button
- Statistics on the number of components is shown on the left
- Larger components are listed with information
- Select a component and click to Save or Add to interface

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 8



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Large Scale Loader: K-Core Analyses

k-degree	Nodes	Components	Link count	Link sum	Density
602	1813	1	968765	968765	0.294729
612	1772	1	943913	943913	0.300611
622	1731	1	918690	918690	0.306602
632	1689	1	892343	892343	0.312804
642	1649	1	866908	866908	0.31881
652	1602	1	836523	836523	0.325951
662	1532	1	790511	790511	0.336814
672	1452	1	737333	737333	0.349728
682	1382	1	690079	690079	0.361312
692	1315	1	644152	644152	0.372509
702	1161	1	536933	536933	0.398342

- Network k-cores can also be investigated
- Click on K-cores and then the Compute button
- Each k-core detected is listed
- Note that k-cores are subsets of each other
- Density increases as K increases
- Largest k-core is k=702 with 1161 nodes!
- Select k-cores to save

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 9

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Large-Scale Techniques

- The following slides are techniques for trimming data within ORA's Network Editor using
 - Node degree
 - Components
 - K-core

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 10



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node Pruning by Degree

- Load the Snowball data which is Twitter data about NATO and Russia in the Baltic region.
- There are 12020 hashtags
- There are 13450 agents
- Select the Agent x Hashtag network
- Click the **Binary link values** checkbox

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 11

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node Pruning by Degree...

	OrDiBAl...	BRBiquz	Marsed6	
1001ptpPL	1	1	2	
adgpb	0	0	1	
Tyrel	0	0	0	
WozzyVA...	0	0	0	
Stanislaw...	0	0	0	
annakubist	0	0	0	
frang_M...	0	0	0	
YourAnsh...	0	0	0	
Tomasz...	0	0	0	
DavidCe...	0	0	0	
zdrapkae	0	0	0	
chiefrobs...	0	0	0	
Ania_PL_DE	0	0	0	
anecrakad7	0	0	0	
Newswi...	0	0	0	
dzejjel	0	0	1	
Zweetab...	0	0	0	
kwirRobert	0	0	0	
warmian...	0	0	0	
szczuka...	0	0	0	

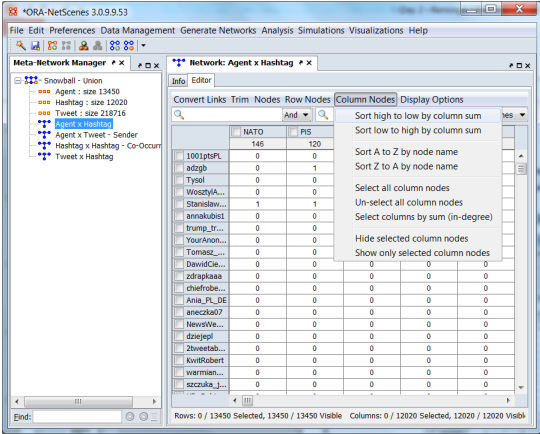
- Use Display Options to show row and column sums
- The column sums give the number of distinct agents that use hashtag
- "Distinct Agents" because we made link weights binary

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 12



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node Pruning by Degree...

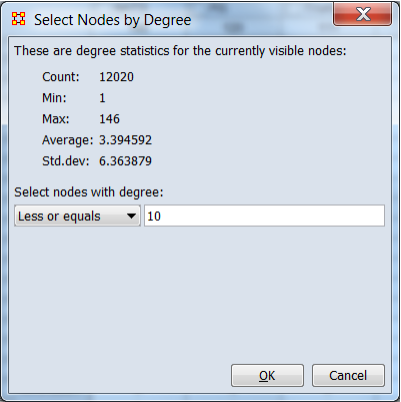


- Sort the columns from high to low by column sum
- Note that NATO, PiS, Trump, Russia, Poland are used the most
- We will reduce the data size by removing hashtags not used by many agents

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 13

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node Pruning by Degree...



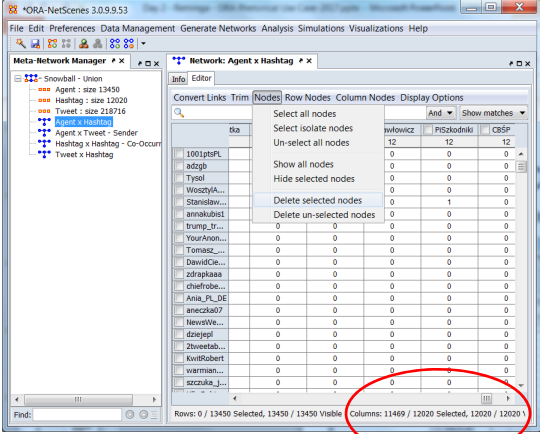
- We will reduce the data size by removing hashtags not used by many agents
- Use the Column Nodes \ Select columns by sum (in-degree)
- Dialog appears showing the distribution of agents using hashtags
- Min is 1
- Min + Stddev \approx 10
- Select these "low" degree hashtags

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 14



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node Pruning by Degree...

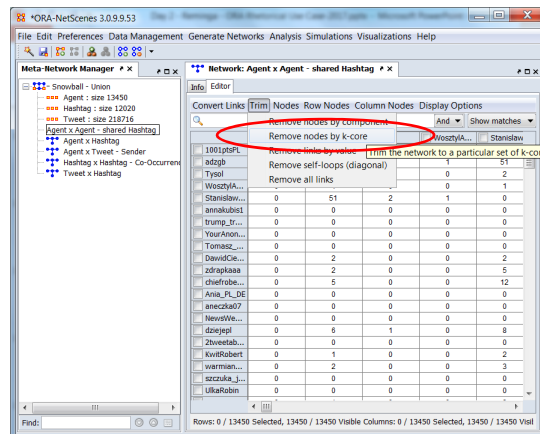


- There are 11469 columns (that is, hashtags) selected
- These all are used by 10 or fewer agents
- Use Nodes \ Delete selected nodes to remove the hashtags
- 551 hashtags remain

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 15

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node pruning by K-Core



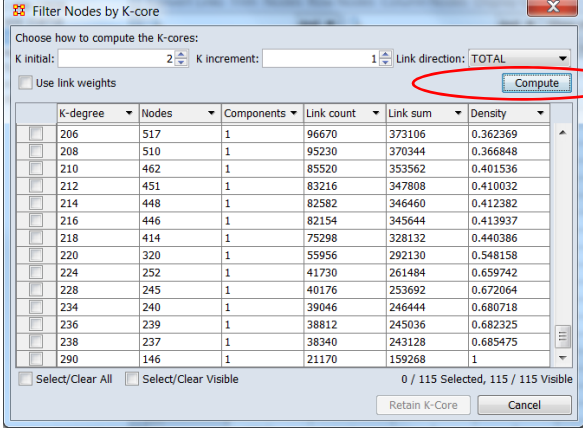
- Suppose we want to find the super-connected groups within the Agent x Agent – shared hashtags
- These are agents talking about the same things
- Use the menu Trim \ Remove nodes by k-core

CASOS June 2020 © 2020 CASOS, Director Kathleen M. Carley 16



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node pruning by K-Core...

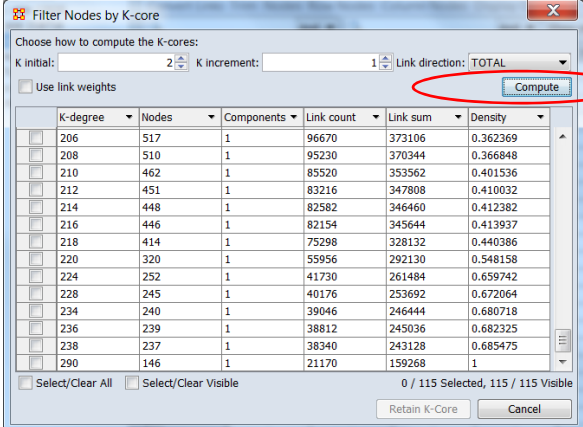


- Click compute on the Filter Nodes by K-core dialog
- The K-cores are computed
- A K-core contains nodes that all have degree $\geq k$
- K-cores are subsets of each other
- The larger the K, the more shared hashtags
- Our largest core has 290 fully connected nodes! (density = 1)

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 17

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node pruning by K-Core...



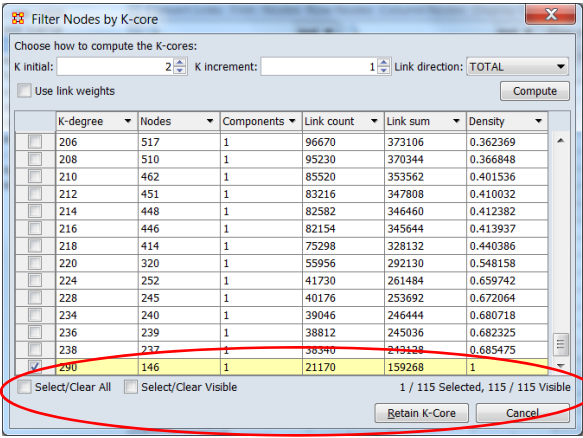
- Click compute on the Filter Nodes by K-core dialog
- The K-cores are computed
- A K-core contains nodes that all have degree $\geq k$
- K-cores are subsets of each other
- The larger the K, the more shared hashtags
- Our largest core has 290 fully connected nodes! (density = 1)

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 18



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node pruning by K-Core...

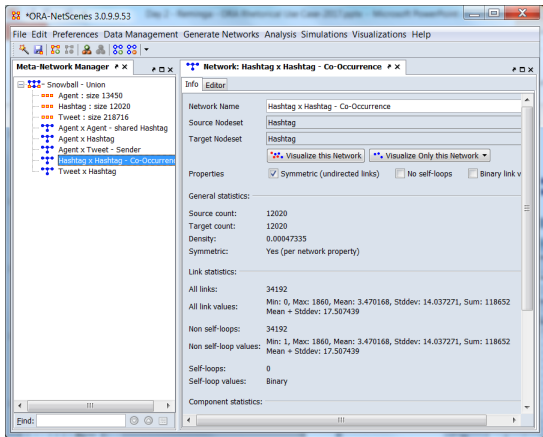


- Select a k-core
- The largest k-selected determines what will be kept
- Click Retain K-core to reduce the agents to this core agent group

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 19

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Node pruning by K-Core...



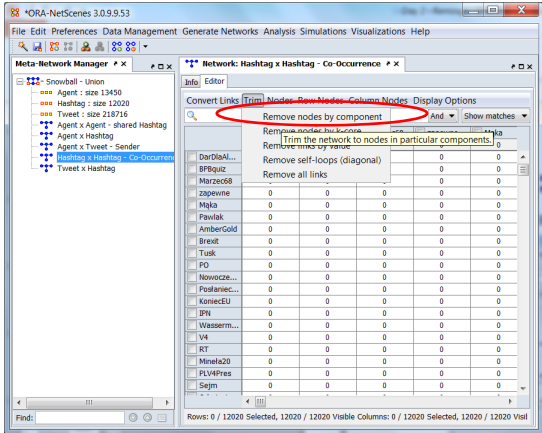
- We could also remove links from the Semantic Network (concept x concept)
- Usually trim out weaker links
- Select the Hashtag x Hashtag network

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 20



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Prune by Component

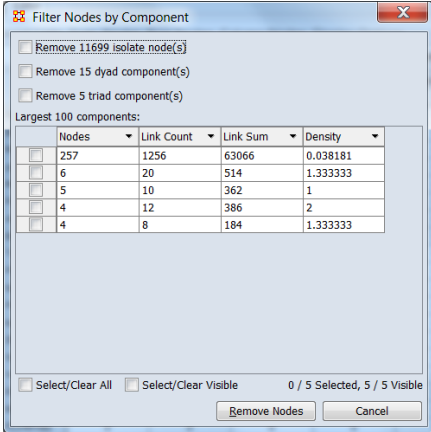


- Click on the network editor
- Use the Trim \ Remove nodes by component to view the components in the network

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 21

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Prune by Component...



- We created many isolates
- 15 dyads
- 5 triads
- We can select which components we want to remove and then click the Remove Nodes button

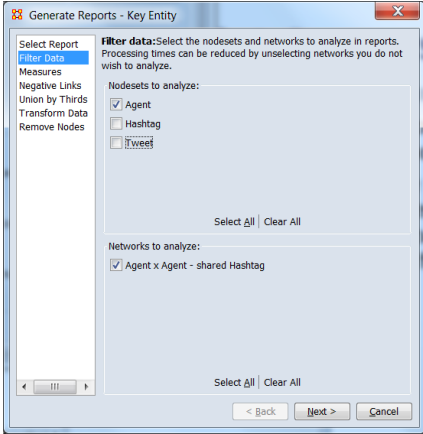
Nodes	Link Count	Link Sum	Density	
<input type="checkbox"/>	257	1256	63066	0.038181
<input type="checkbox"/>	6	20	514	1.333333
<input type="checkbox"/>	5	10	362	1
<input type="checkbox"/>	4	12	386	2
<input type="checkbox"/>	4	8	184	1.333333

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 22



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Select Data For Reports

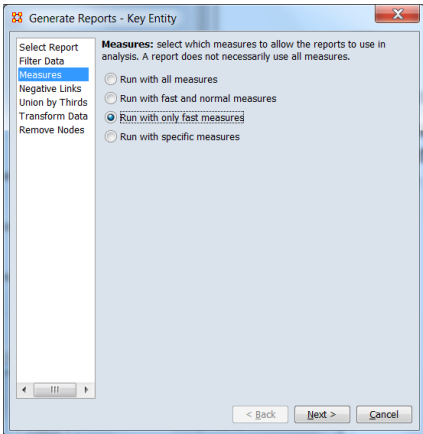


- Goal: only compute measures on networks we are interested in → same time and space
- Suppose we want only to analyze the Agent x Agent – shared hashtags network
- Use the Filter Data tab on the left-hand side
- Then select to analyze only the agent nodeset and its network

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 23

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Select Measures For Reports



- Goal: only run measures that scale well, or that we are interested in
- Click on the Measures tab in the left-hand pane
- Select a category of measures by speed
- Or click on Run with specific measures to fully choose

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 24



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

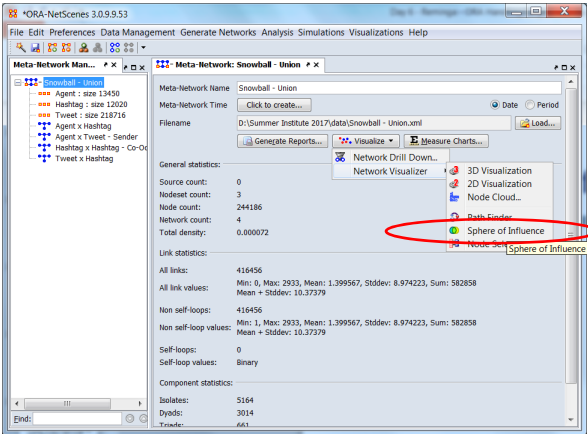
Visualization for Large Scale

- ORA can load data into memory that is too large to visualize
- Rather than visualize a large meta-network in its entirety, one can choose one or more nodes and visualize the sphere of influence or shortest path for the nodes
- This computes the sphere of influence (or shortest path) and then brings the resulting subset of the meta-network into the visualizer

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 25

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Visualization for Large Scale...



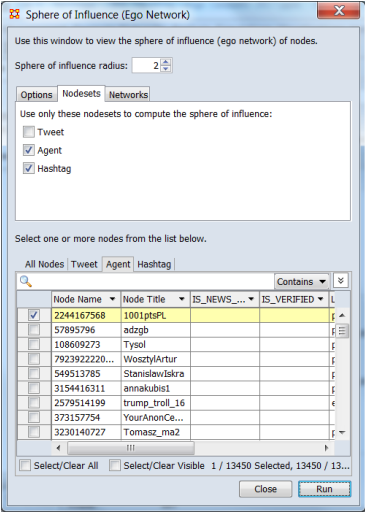
- Load the Day 2 Snowball union dataset
- Select the meta-network
- Click on the Visualize button
- Choose the Sphere of Influence

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 26



Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Visualization for Large Scale...

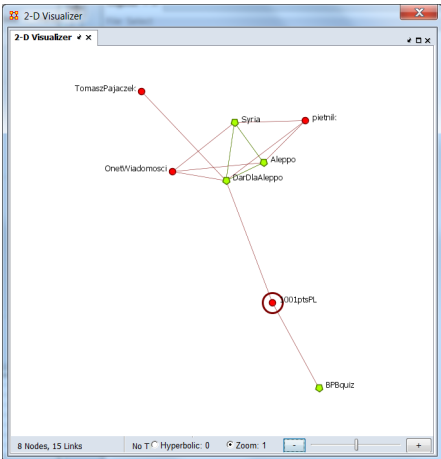


- Choose to visualize only Agents and Hashtags (ignore tweets)
- Use a radius of 2 to show links 1 and 2 steps out
- This is another optimization – do not show the tweets that underlie the Agents and Hashtags)
- Click on the first agent

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 27

Carnegie Mellon
IST Institute for SOFTWARE RESEARCH

Visualization for Large Scale...



- The resulting network is loaded into the visualizer
- The sphere of influence embedded within the large meta-network was computed and only the result brought into the visualizer

CASOS
June 2020 © 2020 CASOS, Director Kathleen M. Carley 28

