Center for Public Health Systems Science

GEORGE WARREN BROWN
SCHOOL OF SOCIAL WORK

Washington University in St.Louis

Network Analysis: Data Wrangling for Evaluators Familiar with SNA

Bobbi J. Carothers

American Evaluation Association
Coffee Break Webinar
8/27/2015

Steps to a Successful Network Analysis

1. Decide who is in the network
2. Decide on network measurements
3. Collect your data
4. Manage your data
5. Analyze your data
6. Visualize your data

Step 1: Network Boundary – Who is in the network?

Who to Include?

- Be guided by the relationships you want to measure
- Laumann criteria
  - Positional: formal membership
  - Reputational: knowledgeable person names members
  - Event: participation in activity of interest
  - Relational: contact with others in the network

Laumann, Mardsen, & Prensky (1983). Boundary specification problem in network analysis.

Step 2: Network Measurements – What relationships are you interested in?

Possible Relationships

- Guided by purpose of the initiative you're evaluating
- Whatever people/organizations are *doing together*
  - Publication co-authorship
  - Amount of contact
  - Level of collaboration
  - Partnership activities
  - Flow of resources (money, information)
  - Satisfaction with communication, collaboration, mentoring, etc.
  - Barriers experienced with partners
  - Dissemination

Step 3: Data Collection – How can you obtain information about relationships?

Online Survey

- Network-specific tools

Tool	URL	Notes
Network Genie	https://secure.networkgenie.com/	Pay
ONASurveys	https://www.s2.onasurveys.com/	Pay
Partner Tool	http://www.partnertool.net/	Pay
OpenEddi	http://www.openeddi.com/	In development, free?
Egoweb	https://github.com/qualintitative/egoweb	In development, free
Polinode	https://www.polinode.com/	Pay

- General online survey platforms
  - Anything that allows display logic and text piped in from responses will work
  - SurveyMonkey (paid)
  - REDCap
  - Qualtrics
- Formats
  - Free recall
  - Roster

Format: Free Recall

- Start with 1 or 2 *name generator* questions asking participants to list who they are connected to or aware of in the network
- Use the piped text feature of the online survey tool to display participant-generated names in subsequent network questions
- Benefits
  - Can “snowball” participants beyond original delineation
- Drawbacks
  - Cleaning creative spelling
  - Participants may be uncomfortable/unwilling to name partners
  - Recalling names → high participant burden
  - Contacting snowballed names → high researcher burden

Format: Roster

- Present participant with a full list of network partners to answer about
- Benefits
  - Easy to clean & manage data
  - Easier for participants to recognize names than to recall them
- Drawbacks
  - Not feasible with very large networks
  - Comprehensive delineation essential

	Yes	No
John Smith		
Tom Parker		
Etc…		

Roster Tips

- Start with a screening question to filter out non-connected partners in later questions (online survey display logic)

	Yes	No
John Smith		
Tom Parker		
Etc…		

- Order of names on roster questions = order of participant IDs
  - Data will export in an N x N matrix
  - Aids in later data management

	John Smith	Tom Parker
John Smith		
Tom Parker		

Step 4: Data Management – How do you get network analysis programs to read your data?

Free recall vs. Roster formats

Data Management Goal

- Most network analysis programs can read files derived from an
  - Arc list

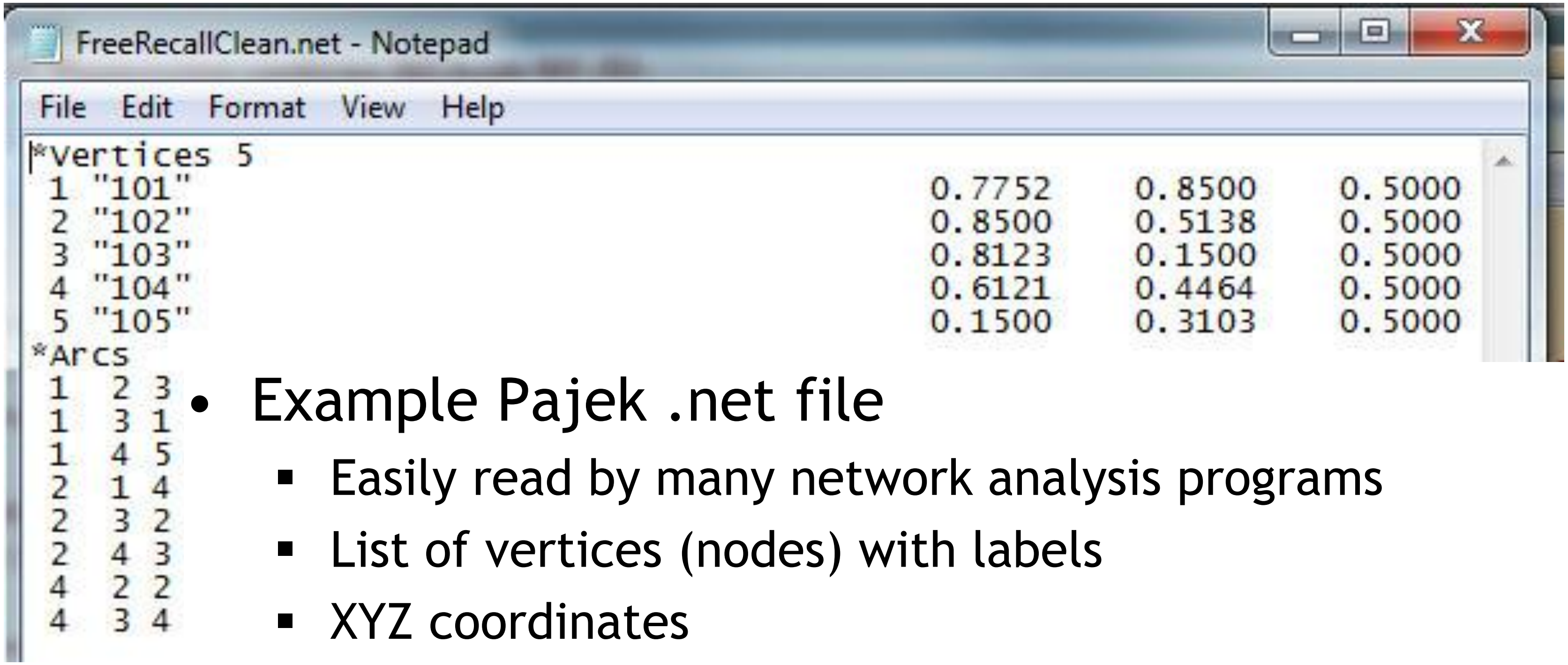
From	To	Value
John Smith	Tom Parker	3
John Smith	Tina Jones	5
Tom Parker	John Smith	4
Tina Jones	Tom Parker	2

or

  - N X N matrix

	John Smith	Tom Parker	Tina Jones
John Smith		3	5
Tom Parker	4		
Tina Jones		2	

Result to Aim For



```
*Vertices 5
 1 "101"                                          0.7752    0.8500    0.5000
 2 "102"                                          0.8500    0.5138    0.5000
 3 "103"                                          0.8123    0.1500    0.5000
 4 "104"                                          0.6121    0.4464    0.5000
 5 "105"                                          0.1500    0.3103    0.5000
*Arcs
 1   2 3
 1   3 1
 1   4 5
 2   1 4
 2   3 2
 2   4 3
 4   2 2
 4   3 4
```

- Example Pajek .net file
  - Easily read by many network analysis programs
  - List of vertices (nodes) with labels
  - XYZ coordinates
  - List of arcs (directional) or edges (non-directional)
    - From
    - To
    - Value (if applicable)

Handy Tools

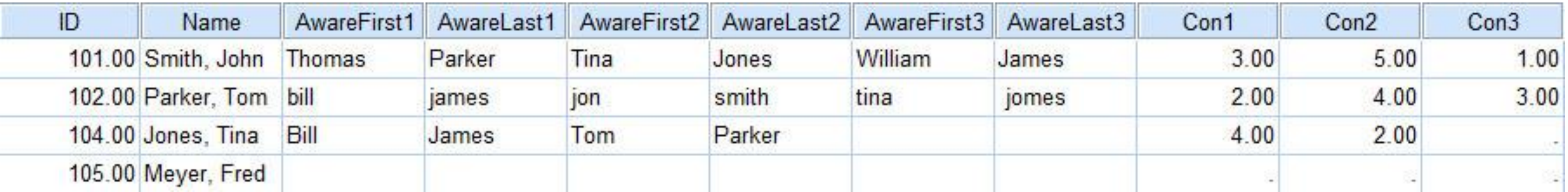
- Pajek (pronounced “pie-yack,” Slovene for “spider”)
  - Network analysis software
  - Useful for fine-tuning network data & performing analyses
  - http://pajek.imfm.si/doku.php?id=pajek
  - Free!
- txt2pajek
  - Turns arc lists into Pajek .net files
  - http://www.pfeffer.at/txt2pajek/
  - Free!
- UCINet
  - Network analysis software, useful for converting matrix files to .net files, sorting .net files
  - https://sites.google.com/site/ucinetsoftware/home
  - Students: $40, Faculty & Government: $150, Others: $250
- Excel, SPSS/SAS/Stata

Data Management Tips

- Convert partner names to numeric IDs with a uniform number of digits
    - 101, 102, 103, etc.
    - Some programs don’t recognize leading zeros (001, 002)
    - Some programs will otherwise sort like this: 1, 10, 11, 2, 21, 22… etc.
    - Different programs may not sort text strings consistently due to different handling of spaces and capitalizations
- Important to match order of network data with order of attribute data

Free Recall Format: Raw Data

- Data will look something like this:

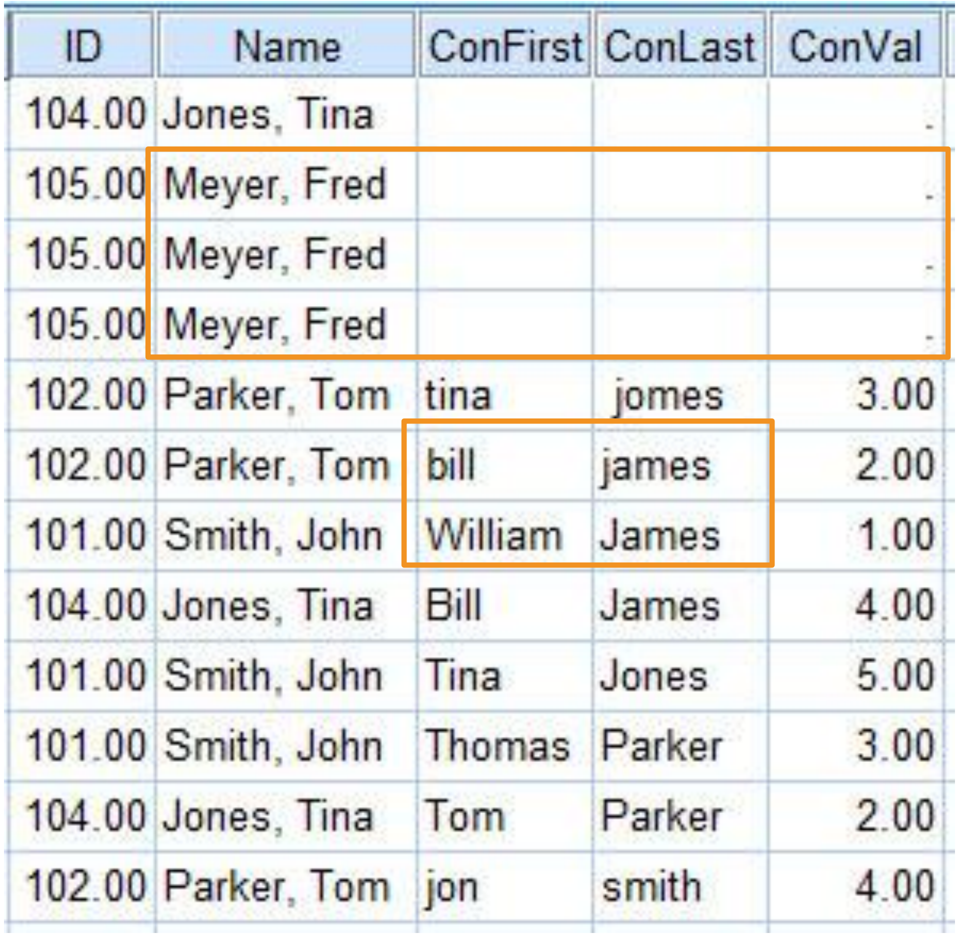


ID	Name	AwareFirst1	AwareLast1	AwareFirst2	AwareLast2	AwareFirst3	AwareLast3	Con1	Con2	Con3
101.00	Smith, John	Thomas	Parker	Tina	Jones	William	James	3.00	5.00	1.00
102.00	Parker, Tom	bill	james	jon	smith	tina	jomes	2.00	4.00	3.00
104.00	Jones, Tina	Bill	James	Tom	Parker			4.00	2.00	.
105.00	Meyer, Fred							.	.	.

- Elements
  - Participant ID and Name, sorted by ID
  - First and last names of people participants listed in awareness name generator
  - Value for the level of contact for each partner
  - Some participants may not have nominated partners
- Strategy: create an arc list that can be converted to a .net file by txt2pajek

Free Recall Format: Transformation

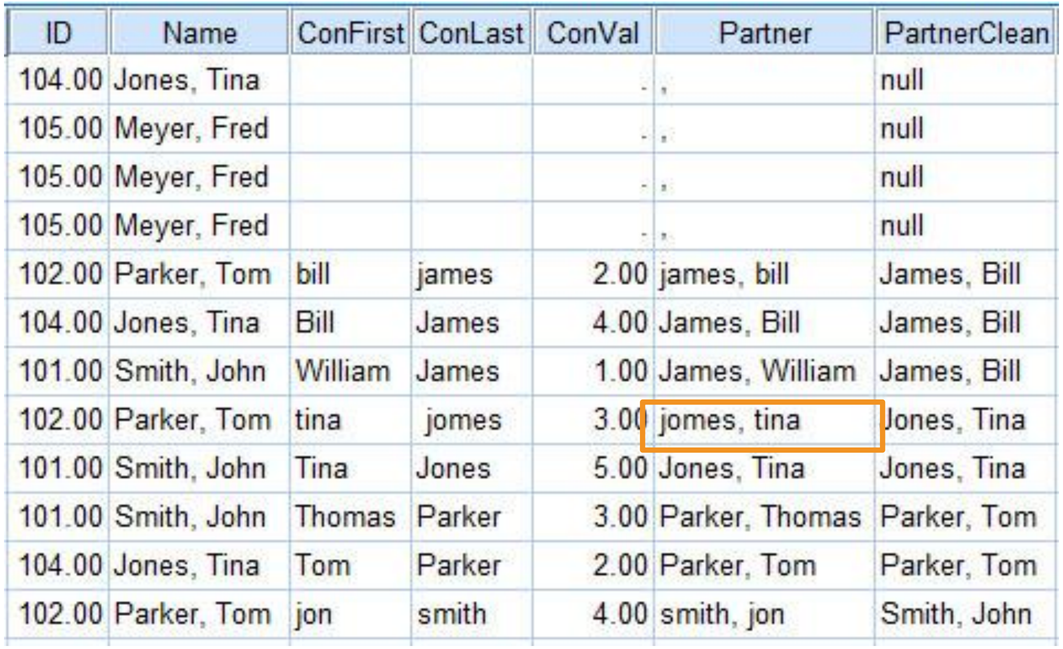
- Convert to a rough arc list
  - Single columns for
    - Fist name
    - Last name
    - Contact value
  - Commands
    - SPSS: varstocases
    - SAS: proc transpose?
    - Stata: reshape long
  - Be sure to retain cases even when partner information is blank (isolate)
  - Sort by last name of nominated partners



ID	Name	ConFirst	ConLast	ConVal
104.00	Jones, Tina			.
105.00	Meyer, Fred			.
105.00	Meyer, Fred			.
105.00	Meyer, Fred			.
102.00	Parker, Tom	tina	jomes	3.00
102.00	Parker, Tom	bill	james	2.00
101.00	Smith, John	William	James	1.00
104.00	Jones, Tina	Bill	James	4.00
101.00	Smith, John	Tina	Jones	5.00
101.00	Smith, John	Thomas	Parker	3.00
104.00	Jones, Tina	Tom	Parker	2.00
102.00	Parker, Tom	jon	smith	4.00

Free Recall Format: Clean, Clean, Clean

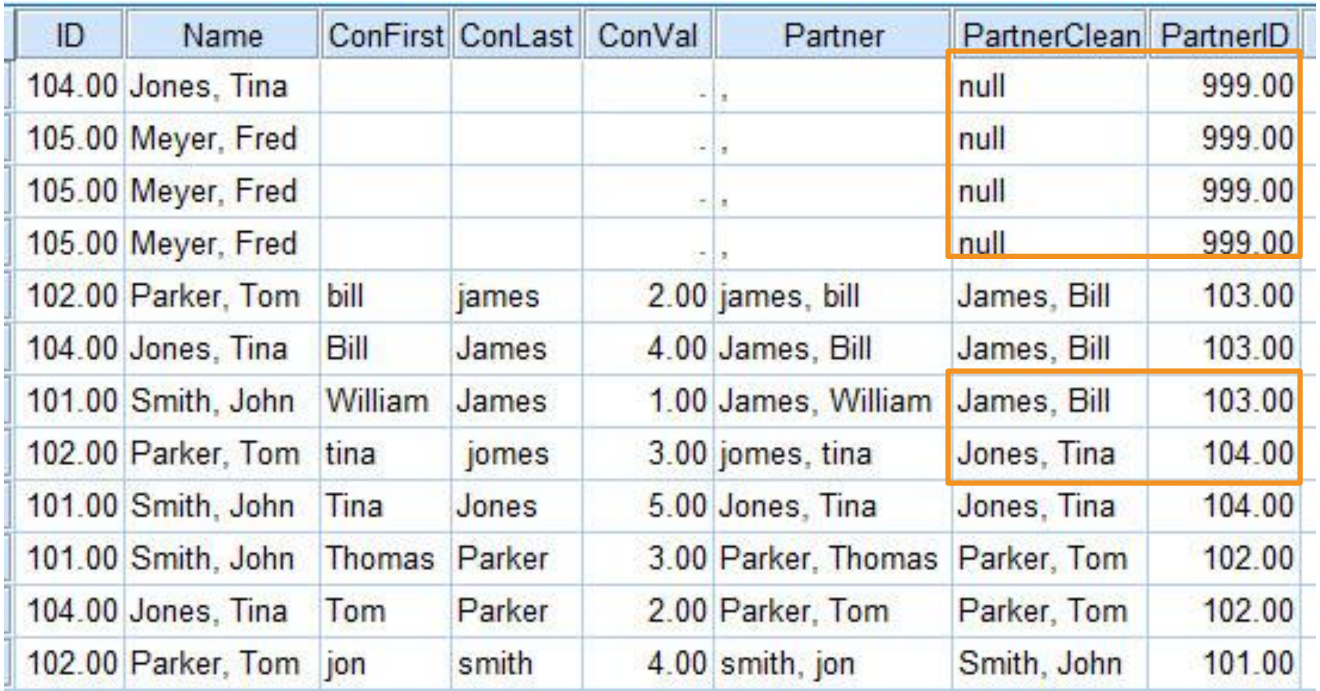
- Clean nominated partner names so they are consistent
  - Concatenate last and first names, trimming extra spaces on the left and right
  - Fix creative spellings and capitalizations (recode)



ID	Name	ConFirst	ConLast	ConVal	Partner	PartnerClean
104.00	Jones, Tina			.	,	null
105.00	Meyer, Fred			.	,	null
105.00	Meyer, Fred			.	,	null
105.00	Meyer, Fred			.	,	null
102.00	Parker, Tom	bill	james	2.00	james, bill	James, Bill
104.00	Jones, Tina	Bill	James	4.00	James, Bill	James, Bill
101.00	Smith, John	William	James	1.00	James, William	James, Bill
102.00	Parker, Tom	tina	jomes	3.00	jomes, tina	Jones, Tina
101.00	Smith, John	Tina	Jones	5.00	Jones, Tina	Jones, Tina
101.00	Smith, John	Thomas	Parker	3.00	Parker, Thomas	Parker, Tom
104.00	Jones, Tina	Tom	Parker	2.00	Parker, Tom	Parker, Tom
102.00	Parker, Tom	jon	smith	4.00	smith, jon	Smith, John

Free Recall Format : ID Numbers

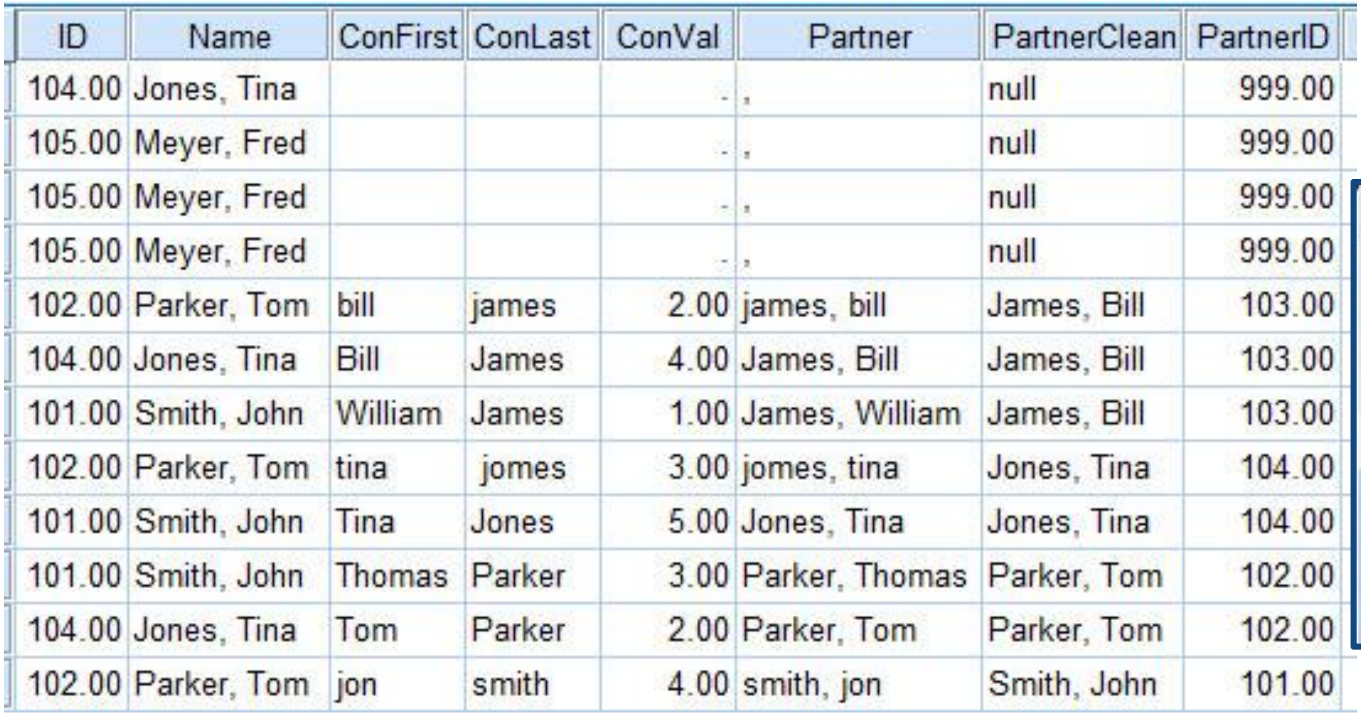
- Assign an ID number to partner names (recode)
  - Match w/ original ID if a participant or part of original delineation
  - Create new ID if not part of original delineation and you want to snowball
  - Add ID for null node



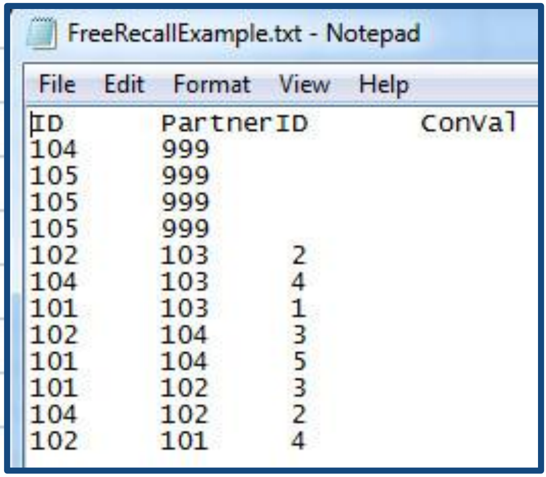
ID	Name	ConFirst	ConLast	ConVal	Partner	PartnerClean	PartnerID
104.00	Jones, Tina			.	,	null	999.00
105.00	Meyer, Fred			.	,	null	999.00
105.00	Meyer, Fred			.	,	null	999.00
105.00	Meyer, Fred			.	,	null	999.00
102.00	Parker, Tom	bill	james	2.00	james, bill	James, Bill	103.00
104.00	Jones, Tina	Bill	James	4.00	James, Bill	James, Bill	103.00
101.00	Smith, John	William	James	1.00	James, William	James, Bill	103.00
102.00	Parker, Tom	tina	jomes	3.00	jomes, tina	Jones, Tina	104.00
101.00	Smith, John	Tina	Jones	5.00	Jones, Tina	Jones, Tina	104.00
101.00	Smith, John	Thomas	Parker	3.00	Parker, Thomas	Parker, Tom	102.00
104.00	Jones, Tina	Tom	Parker	2.00	Parker, Tom	Parker, Tom	102.00
102.00	Parker, Tom	jon	smith	4.00	smith, jon	Smith, John	101.00

Free Recall Format : Arc List

- Save out as tab-delimited text file
  - Keep ID, PartnerID, and value only
  - Variable order is important
- Looks like lower part of Pajek .net file



ID	Name	ConFirst	ConLast	ConVal	Partner	PartnerClean	PartnerID
104.00	Jones, Tina			.	,	null	999.00
105.00	Meyer, Fred			.	,	null	999.00
105.00	Meyer, Fred			.	,	null	999.00
105.00	Meyer, Fred			.	,	null	999.00
102.00	Parker, Tom	bill	james	2.00	james, bill	James, Bill	103.00
104.00	Jones, Tina	Bill	James	4.00	James, Bill	James, Bill	103.00
101.00	Smith, John	William	James	1.00	James, William	James, Bill	103.00
102.00	Parker, Tom	tina	jomes	3.00	jomes, tina	Jones, Tina	104.00
101.00	Smith, John	Tina	Jones	5.00	Jones, Tina	Jones, Tina	104.00
101.00	Smith, John	Thomas	Parker	3.00	Parker, Thomas	Parker, Tom	102.00
104.00	Jones, Tina	Tom	Parker	2.00	Parker, Tom	Parker, Tom	102.00
102.00	Parker, Tom	jon	smith	4.00	smith, jon	Smith, John	101.00

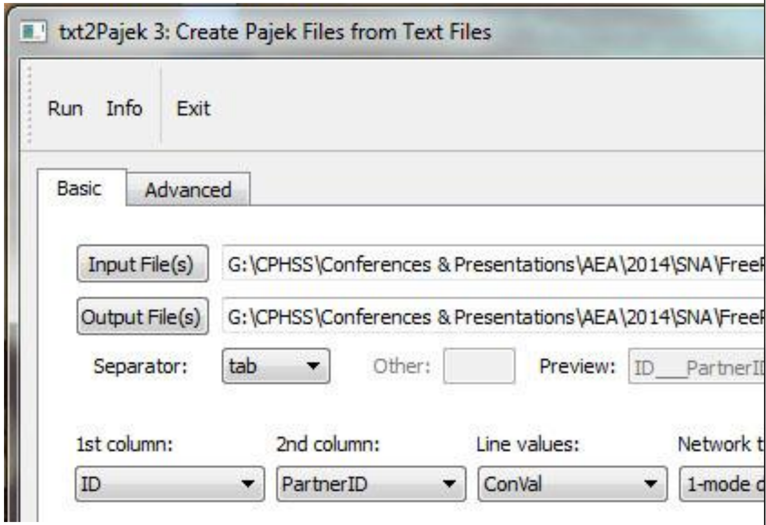


FreeRecallExample.txt - Notepad

```
ID      PartnerID       ConVal
104     999
105     999
105     999
105     999
102     103     2
104     103     4
101     103     1
102     104     3
101     104     5
101     102     3
104     102     2
102     101     4
```

Free Recall Format : Convert to Pajek

- txt2Pajek



txt2Pajek 3: Create Pajek Files from Text Files

Input File(s) G:\CPHSS\Conferences & Presentations\AEA\2014\SNA\FreeF

Output File(s) G:\CPHSS\Conferences & Presentations\AEA\2014\SNA\FreeF

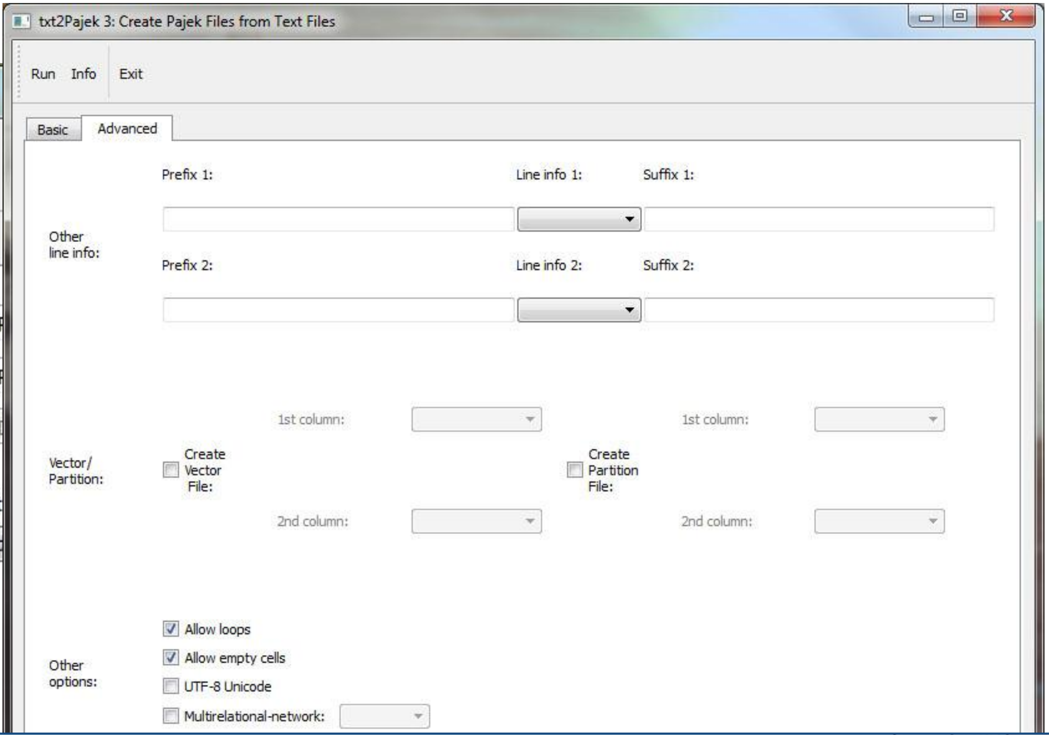
Separator: tab Other: Preview: ID___PartnerII

1st column: ID

2nd column: PartnerID

Line values: ConVal

Network t 1-mode c



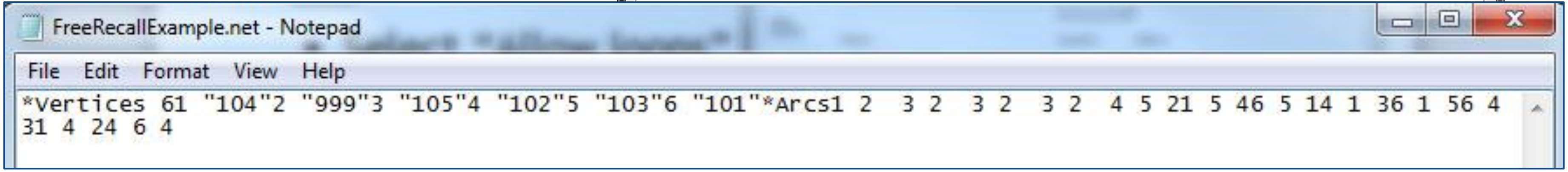
txt2Pajek 3: Create Pajek Files from Text Files

Other line info: Prefix 1: Line info 1: Suffix 1:

Prefix 2: Line info 2: Suffix 2:

Vector/ Partition: □ Create Vector File: 1st column: 2nd column: □ Create Partition File: 1st column: 2nd column:

Other options:
☑ Allow loops
☑ Allow empty cells
□ UTF-8 Unicode
□ Multirelational-network:



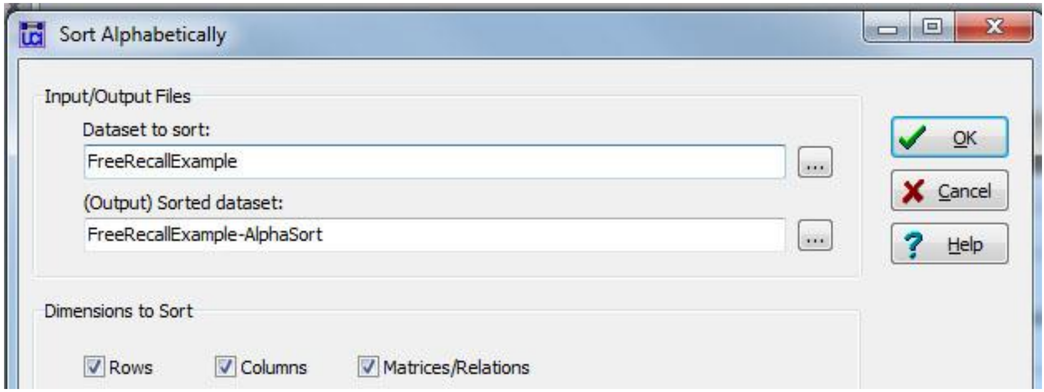
FreeRecallExample.net - Notepad

```
*Vertices 61 "104"2 "999"3 "105"4 "102"5 "103"6 "101"*Arcs1 2  3 2  3 2  3 2  4 5 21 5 46 5 14 1 36 1 56 4
31 4 24 6 4
```

- Hmm, still needs some work

Free Recall Format : Sort Nodes

- Order of nodes in network file = order in attribute files
- UCINet
  - Data → Import text file → Pajek
  - Data → Sort Alphabetically
    - Select non-Crd ##h file
  - Data → Export → Pajek → Network
    - Select AlphaSort version
    - Do not launch Pajek (old version)



Sort Alphabetically

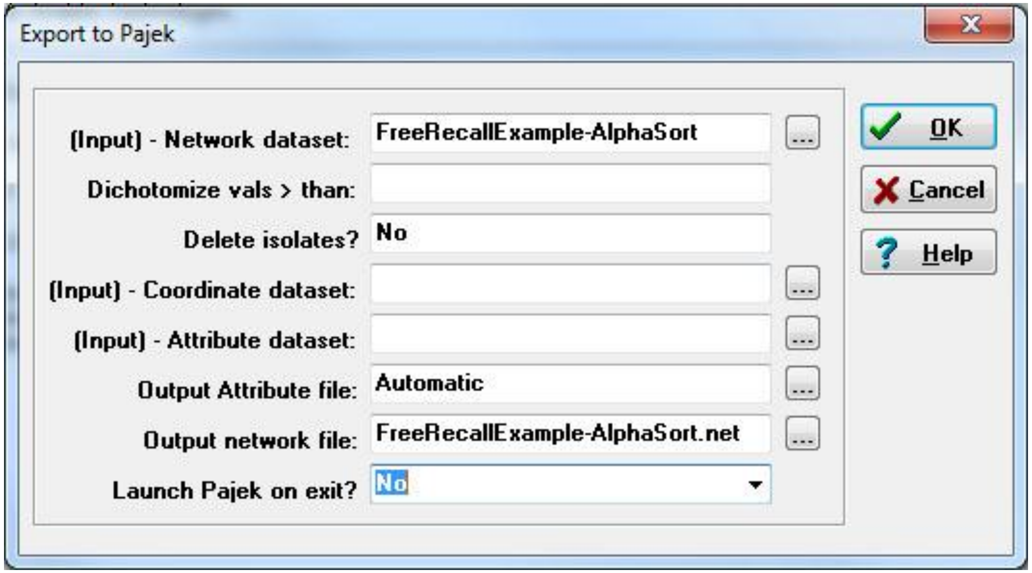
Input/Output Files

Dataset to sort: FreeRecallExample

(Output) Sorted dataset: FreeRecallExample-AlphaSort

Dimensions to Sort

☑ Rows ☑ Columns ☑ Matrices/Relations



Export to Pajek

(Input) - Network dataset: FreeRecallExample-AlphaSort

Dichotomize vals > than:

Delete isolates? No

(Input) - Coordinate dataset:

(Input) - Attribute dataset:

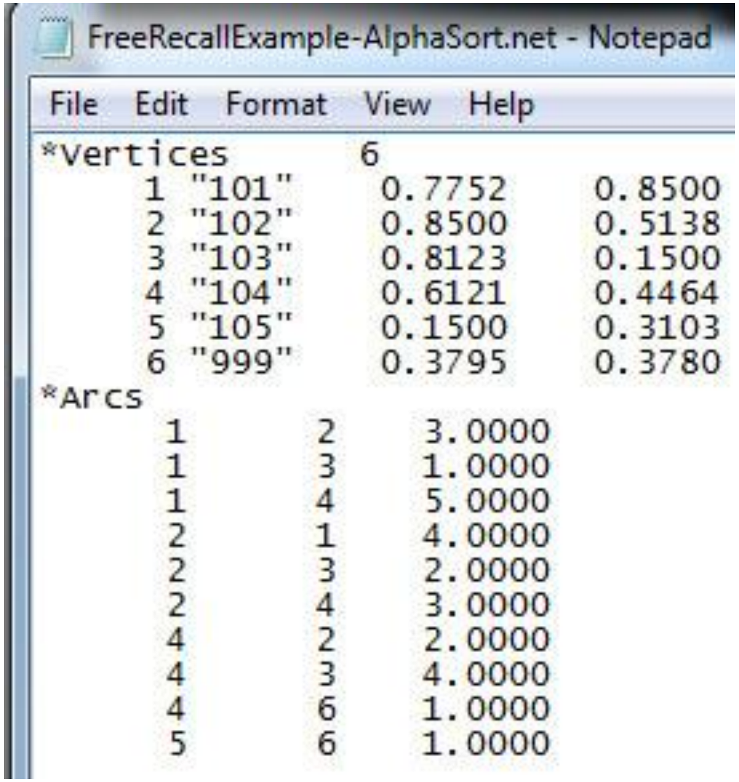
Output Attribute file: Automatic

Output network file: FreeRecallExample-AlphaSort.net

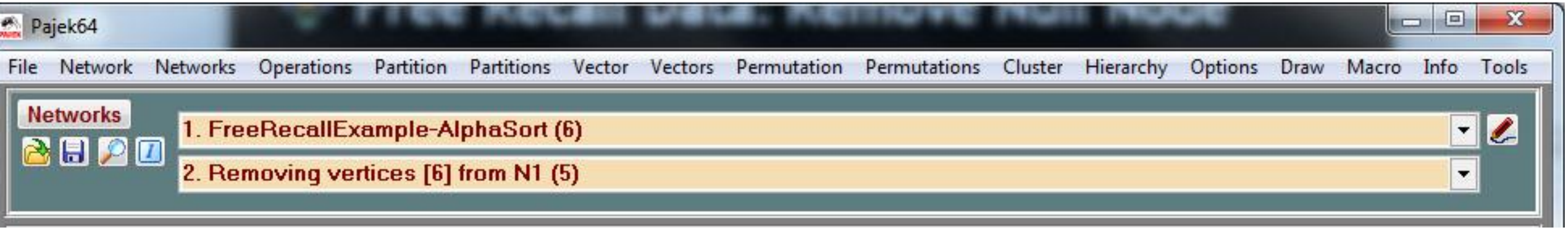
Launch Pajek on exit? No

Free Recall Format : Remove Null Node

- Pajek
  - Drag & drop AlphaSort file into first network box
  - File → Network → Change Label to clean text
  - Network → Create New Network → Transform → Remove → Selected Vertices → enter appropriate # (in this case, 6)


FreeRecallExample-AlphaSort.net - Notepad
File Edit Format View Help

```
*Vertices      6
       1 "101"      0.7752      0.8500
       2 "102"      0.8500      0.5138
       3 "103"      0.8123      0.1500
       4 "104"      0.6121      0.4464
       5 "105"      0.1500      0.3103
       6 "999"      0.3795      0.3780
*Arcs
        1        2      3.0000
        1        3      1.0000
        1        4      5.0000
        2        1      4.0000
        2        3      2.0000
        2        4      3.0000
        4        2      2.0000
        4        3      4.0000
        4        6      1.0000
        5        6      1.0000
```

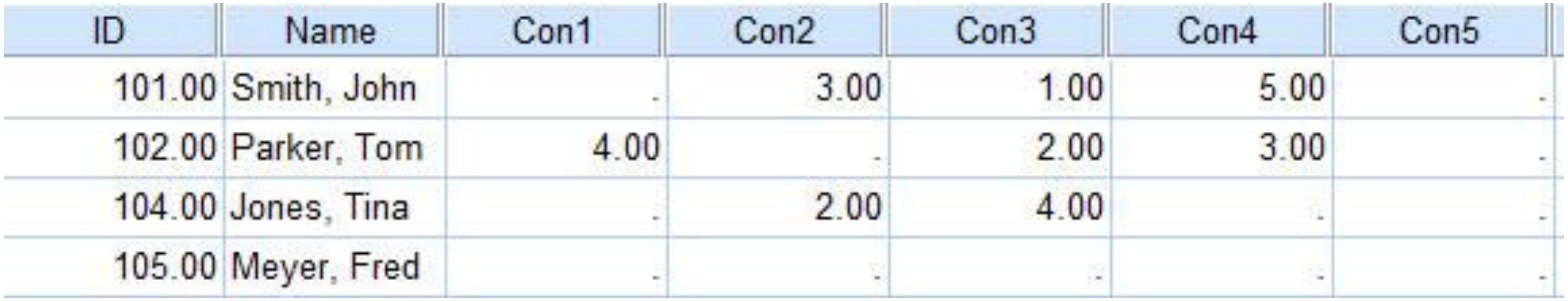

Pajek64
File Network Networks Operations Partition Partitions Vector Vectors Permutation Permutations Cluster Hierarchy Options Draw Macro Info Tools
Networks

1. FreeRecallExample-AlphaSort (6)

2. Removing vertices [6] from N1 (5)

Roster Format : Raw Data

- Data will look something like this:

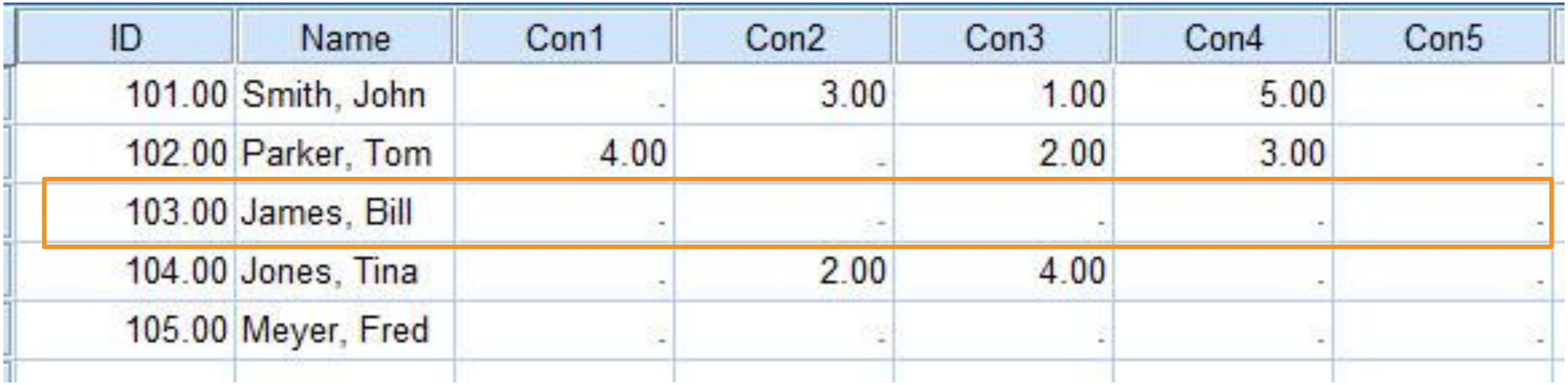


ID	Name	Con1	Con2	Con3	Con4	Con5
101.00	Smith, John	.	3.00	1.00	5.00	.
102.00	Parker, Tom	4.00	.	2.00	3.00	.
104.00	Jones, Tina	.	2.00	4.00	.	.
105.00	Meyer, Fred	.	.	.	.	.

- Elements
  - When sorted by ID, comes close to an N x N matrix
  - Con1 is everyone’s contact rating for John Smith, Con2 is everyone’s contact rating for Tom Parker, etc.
  - “From” is the ID column, “To” is each of the Con columns
- Strategy: create clean N x N matrix, use UCINet to convert to Pajek .net file

Roster Format : Insert Non-Respondents

- Add non-respondents in correct order

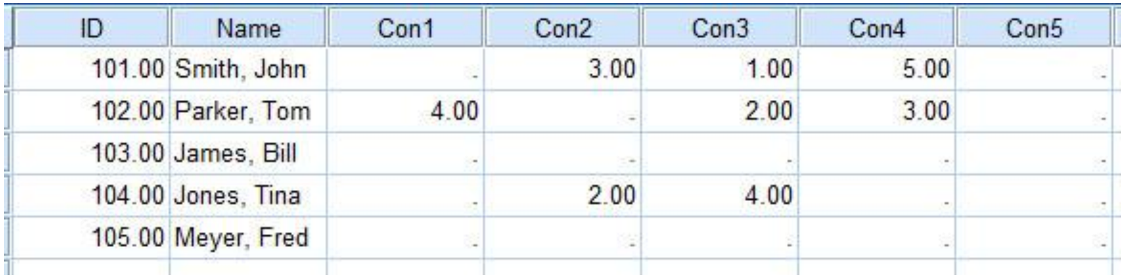


ID	Name	Con1	Con2	Con3	Con4	Con5
101.00	Smith, John	.	3.00	1.00	5.00	.
102.00	Parker, Tom	4.00	.	2.00	3.00	.
103.00	James, Bill	.	.	.	.	.
104.00	Jones, Tina	.	2.00	4.00	.	.
105.00	Meyer, Fred	.	.	.	.	.

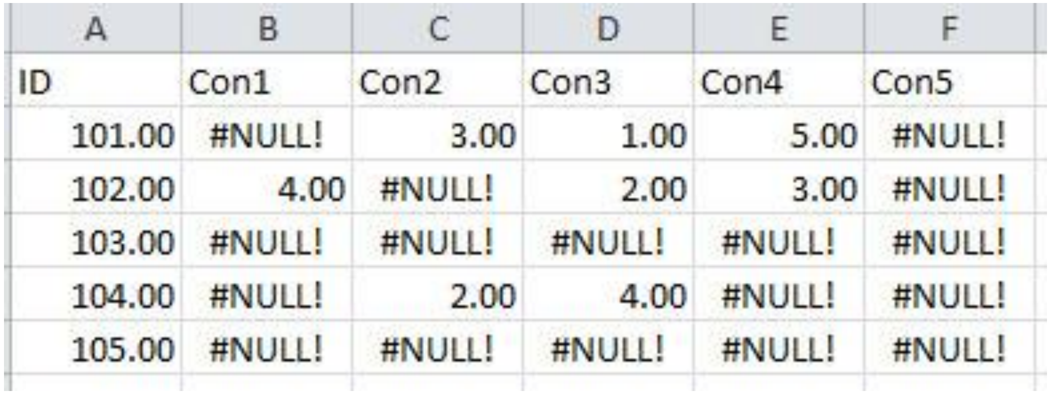
- Aaannnd… that’s it!

Roster Format : Export to Excel

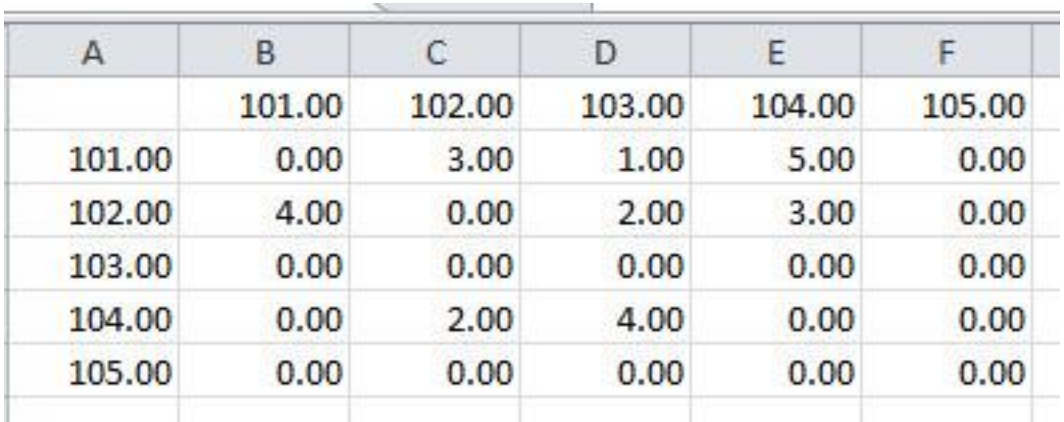
- Export as Excel file (remove Name)
- Clean
  - Clear ID cell
  - Find #NULL! & replace with 0
  - Copy ID numbers and Paste Special → Transpose



ID	Name	Con1	Con2	Con3	Con4	Con5
101.00	Smith, John	.	3.00	1.00	5.00	.
102.00	Parker, Tom	4.00	.	2.00	3.00	.
103.00	James, Bill	.	.	.	.	.
104.00	Jones, Tina	.	2.00	4.00	.	.
105.00	Meyer, Fred	.	.	.	.	.



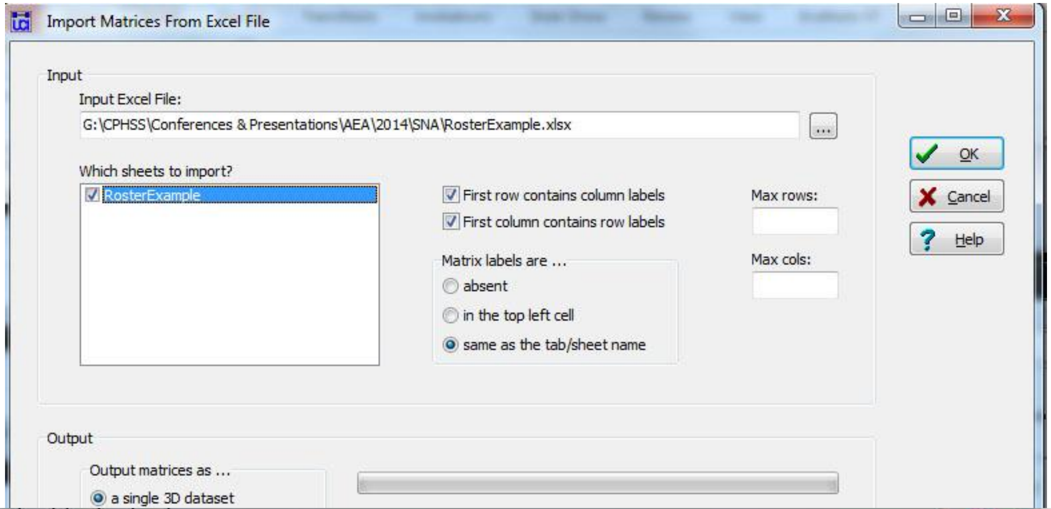
A	B	C	D	E	F
ID	Con1	Con2	Con3	Con4	Con5
101.00	#NULL!	3.00	1.00	5.00	#NULL!
102.00	4.00	#NULL!	2.00	3.00	#NULL!
103.00	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!
104.00	#NULL!	2.00	4.00	#NULL!	#NULL!
105.00	#NULL!	#NULL!	#NULL!	#NULL!	#NULL!



A	B	C	D	E	F
	101.00	102.00	103.00	104.00	105.00
101.00	0.00	3.00	1.00	5.00	0.00
102.00	4.00	0.00	2.00	3.00	0.00
103.00	0.00	0.00	0.00	0.00	0.00
104.00	0.00	2.00	4.00	0.00	0.00
105.00	0.00	0.00	0.00	0.00	0.00

Roster Format : Convert to Pajek

- UCINet
  - Data → Import Excel → Matricies
  - Data → Export → Pajek → Network
    - Do not launch Pajek



Import Matrices From Excel File

Input

Input Excel File: G:\CPHSS\Conferences & Presentations\AEA\2014\SNA\RosterExample.xlsx

Which sheets to import?

☑ RosterExample

☑ First row contains column labels

☑ First column contains row labels

Matrix labels are ...

○ absent

○ in the top left cell

● same as the tab/sheet name

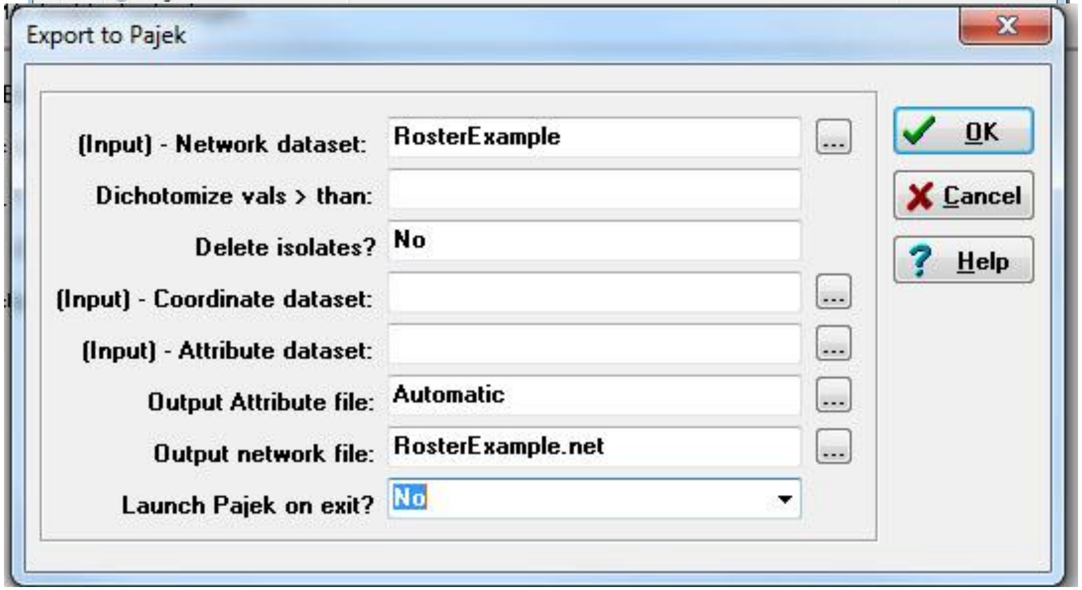
Max rows:

Max cols:

Output

Output matrices as ...

● a single 3D dataset

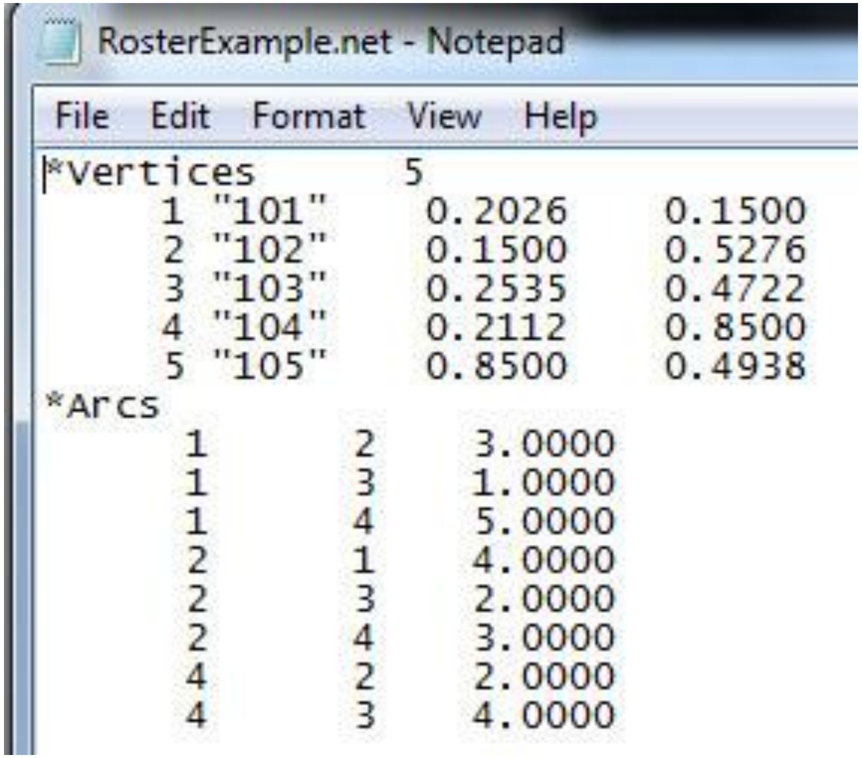


Export to Pajek

(Input) - Network dataset:	RosterExample
Dichotomize vals > than:	
Delete isolates?	No
(Input) - Coordinate dataset:	
(Input) - Attribute dataset:	
Output Attribute file:	Automatic
Output network file:	RosterExample.net
Launch Pajek on exit?	No

Roster Format : Final Product

- Look familiar?



```
*Vertices      5
       1 "101"      0.2026      0.1500
       2 "102"      0.1500      0.5276
       3 "103"      0.2535      0.4722
       4 "104"      0.2112      0.8500
       5 "105"      0.8500      0.4938
*Arcs
        1          2      3.0000
        1          3      1.0000
        1          4      5.0000
        2          1      4.0000
        2          3      2.0000
        2          4      3.0000
        4          2      2.0000
        4          3      4.0000
```

Step 5: Data Analysis – What Is the Structure of the Network?

Network Analysis Software

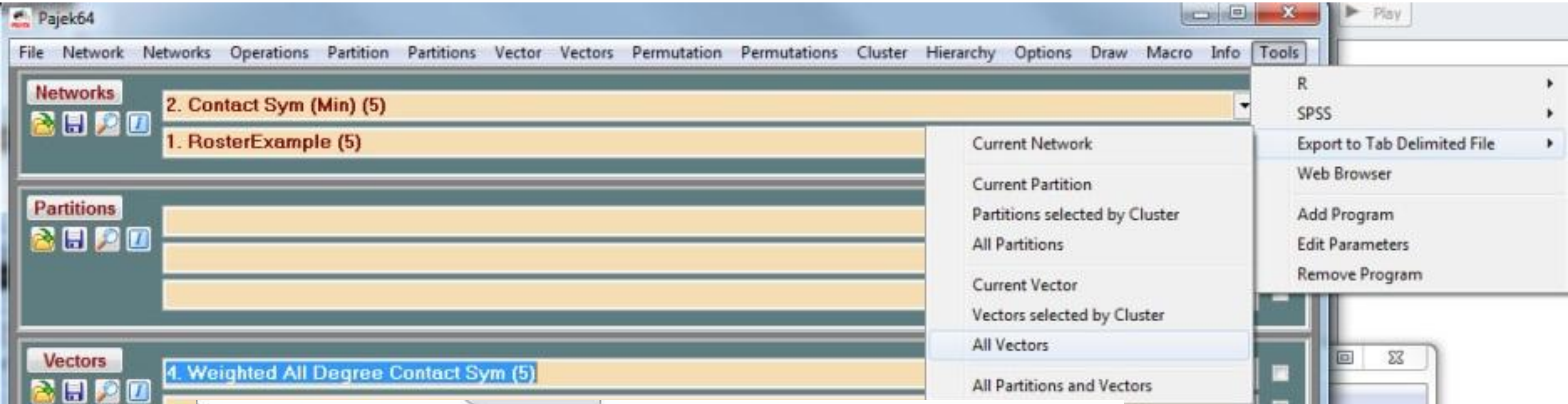
- Pajek
  - http://pajek.imfm.si/doku.php?id=pajek
  - Pros
    - Easy to learn
    - Transparent about what it does
    - Computes many standard network statistics
    - Free!
  - Cons
    - Can be difficult to produce attractive graphics

- Gephi
  - https://gephi.github.io/
  - Pros
    - Easy to learn
    - Easy to produce attractive graphics
    - Free!
  - Cons
    - Less transparent about what it does
    - Computes fewer network statistics
    - Not recently updated, Java incompatibilities

- Strategy
  - Perform analyses in Pajek
  - Transfer numbers to Gephi for visualizations

Exporting Node Network Data

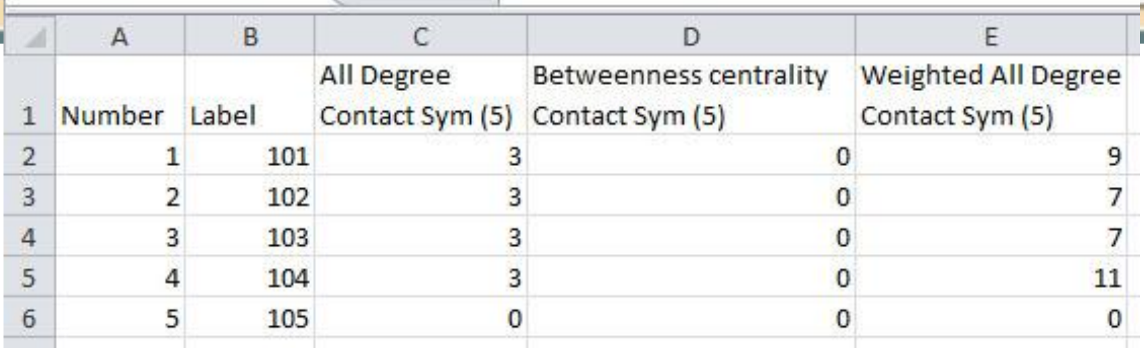
- From Pajek
- Tools → Export to Tab Delimited File → All Vectors (or whichever is most appropriate)



2. Contact Sym (Min) (5)

1. RosterExample (5)

4. Weighted All Degree Contact Sym (5)



	A	B	C	D	E
1	Number	Label	All Degree Contact Sym (5)	Betweenness centrality Contact Sym (5)	Weighted All Degree Contact Sym (5)
2	1	101	3	0	9
3	2	102	3	0	7
4	3	103	3	0	7
5	4	104	3	0	11
6	5	105	0	0	0

Step 6: Network Visualization – What Does the Network Look Like?

or

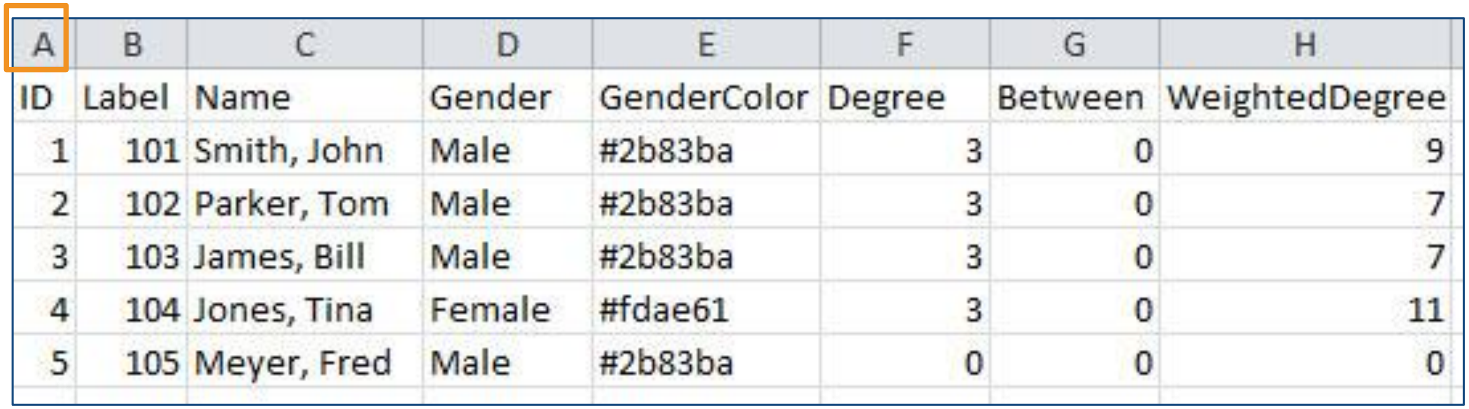
How Do I Make Those Pretty Pictures?

Gephi Resources

- Plugins
  - https://marketplace.gephi.org/
  - Give Color to Nodes: Allows Gephi to read hex color codes
  - Noverlap: Eliminates node overlap
  - SigmaJS: Export interactive networks to the web (install JSON Exporter as well)
  - Many other options available to browse!
- Tutorials
  - http://gephi.github.io/users/
- Fix Java incompatibility:
  - PC: https://forum.gephi.org/viewtopic.php?f=3&t=3580&p=10712#p10712
  - Mac: https://github.com/gephi/gephi/issues/895

Prepare & Export Attribute File to CSV

- Pull attribute and network data from survey and network analysis into one file



A	B	C	D	E	F	G	H
ID	Label	Name	Gender	GenderColor	Degree	Between	WeightedDegree
1	101	Smith, John	Male	#2b83ba	3	0	9
2	102	Parker, Tom	Male	#2b83ba	3	0	7
3	103	James, Bill	Male	#2b83ba	3	0	7
4	104	Jones, Tina	Female	#fdae61	3	0	11
5	105	Meyer, Fred	Male	#2b83ba	0	0	0

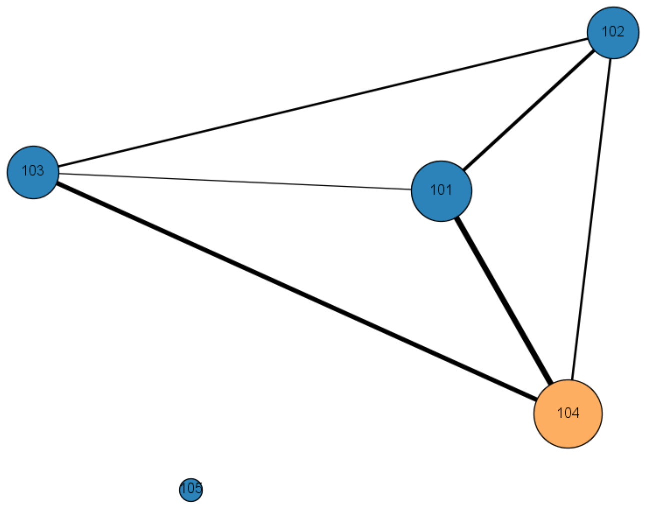
- Change "Number" to "ID" if you're planning to use Gephi for visualizations; ID should be first column
- Gephi can only interpret one color variable at a time if using Give Color to Nodes

Import Data to Gephi

- Import clean .net file
- Import attribute data
  - Data Laboratory section
  - Import Settings: change numeric variables from “String” to “Big Decimal” or “Integer” to allow node sizing

Export Graphic

- SVG, PDF, or PNG options
- If you have Adobe Illustrator, saving to SGV will allow further fine-tuning



Questions?

Bobbi Carothers
bcarothers@wustl.edu

http://cphss.wustl.edu
@cphsswustl