



Intro to Humanities Data:

The Path to Complex Visualization & Statistics

Mia Partlow, @mia_partlow

Digital Humanities Fellow, Information & Library Science
Digital Methods Specialist, Institute for Digital Arts &
Humanities
mapartlo@indiana.edu

Kalani Craig, @kalanicraig

Co-Director, Institute for Digital Arts & Humanities
Clinical Assistant Professor, Department of History
craigkl@Indiana.edu

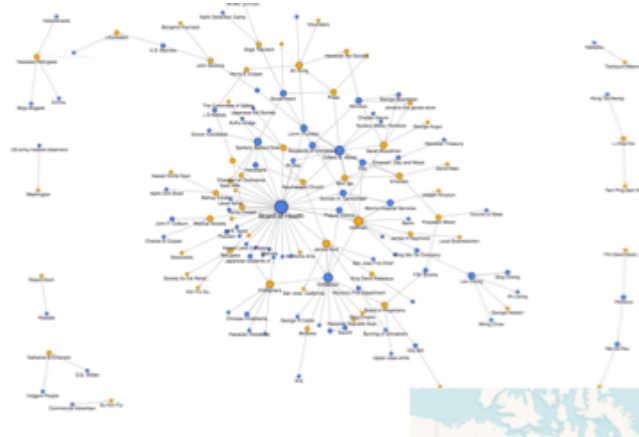
Humanities Data & Statistics

- Humanities Data as *capta* -- “that which has been gathered or created”
 - An active process driven by your goals and tools
 - *Capta* is ambiguous and interpretive
- Data requires cleaning:
 - Creating consistent structures and data types
 - Formatting for your tool or analysis
- Consider what you have or can gather
 - What is the sampling frame? The sample?
 - What type of data do you have?

Making decisions about visualization vs statistics or both

What is your question?

- Network relationships
 - How strongly associated are clusters of interactions?
 - Issues: capta, ego networks
- Geographic relationships
 - How strong is the association between place and subject?
 - Issues: rigid structure, raw data, sampling

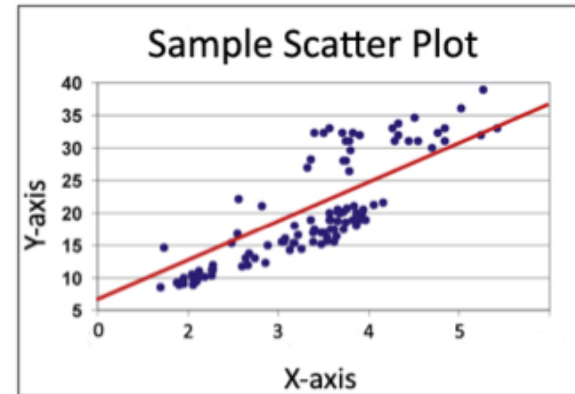


**Relationships are easy to see
but hard to prove**

Making decisions about visualization vs statistics or both

- Regression Analysis: correlation/prediction between two variables
 - **Correlation:** Does use of **#NetNeutrality** hashtag correlate to appearance of **Ajit Pai**'s name in Tweets?
 - **Prediction:** Can existing data on the relationship between two variables help us fill in gaps?
- Chi-Square: relationship between two categorical variables
 - **Correlation:** Are authors from certain countries more likely to write certain genres of literature?
 - **Categories** must be meaningful.

Statistical questions are sometimes prompted by visualizations, and visualizations can be complemented with statistics.



Gestalt psychology and visualization

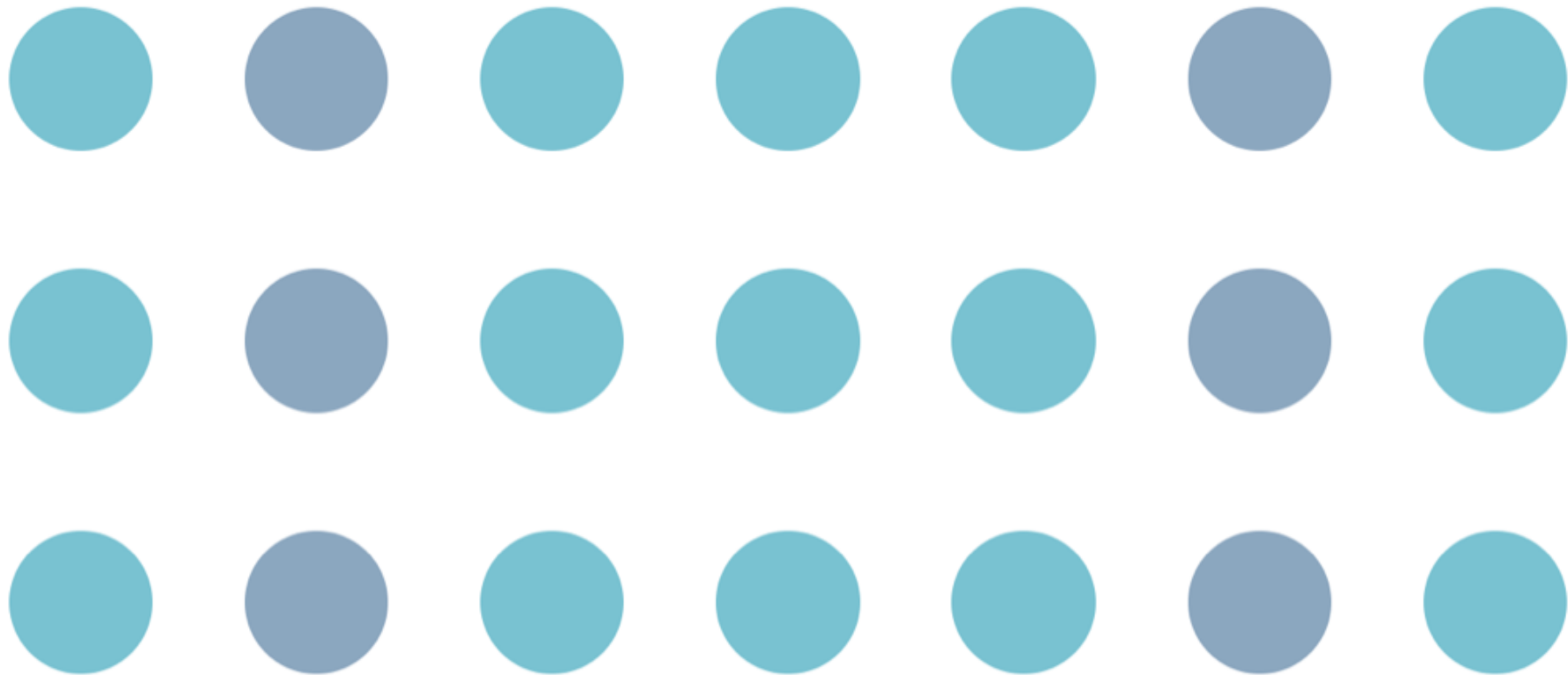
Gestalt psychology

understanding how humans perceive patterns.

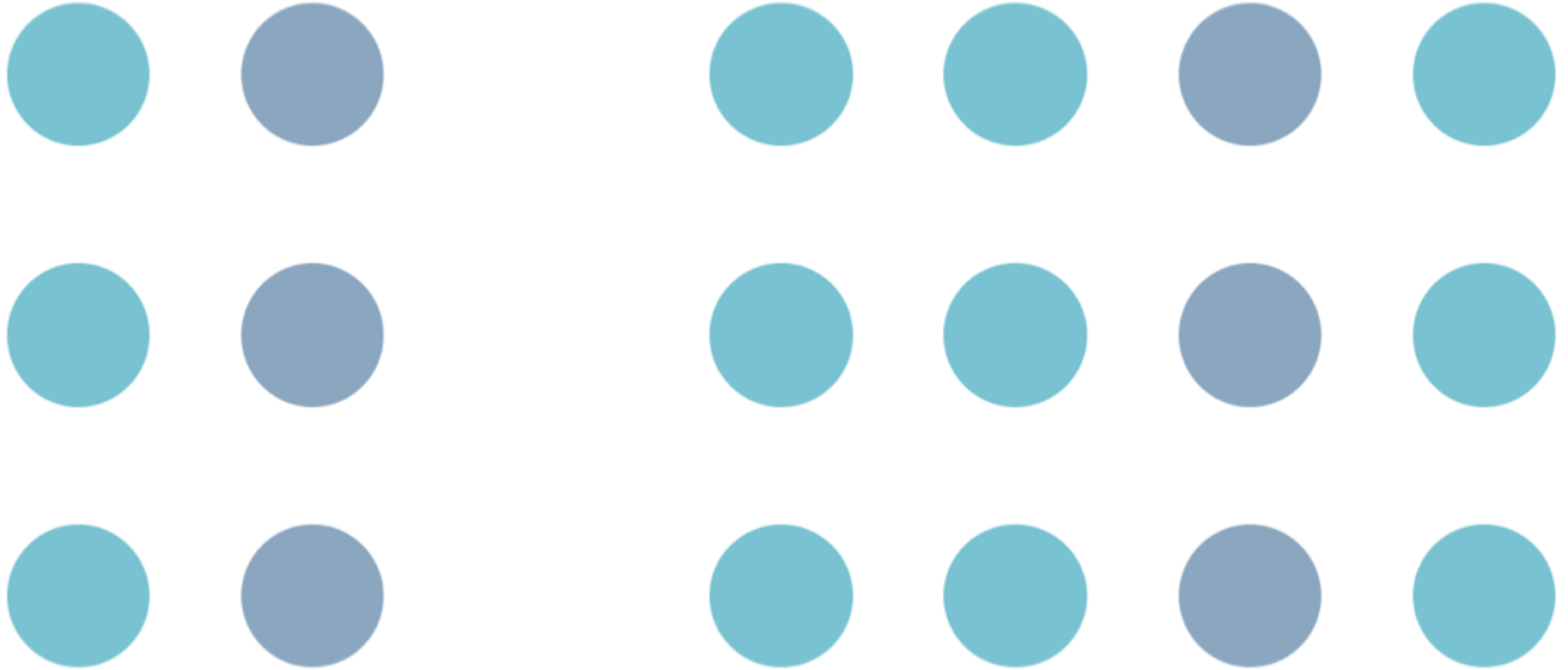
Gestalt psychology in visualization

remembering that we view separate elements as part of a whole pattern

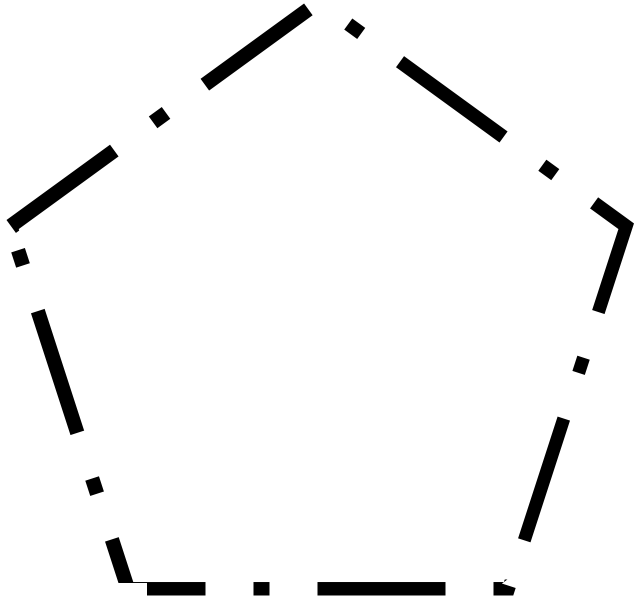
Visual Similarity = Object Similarity



Visual Proximity = Inherent Grouping

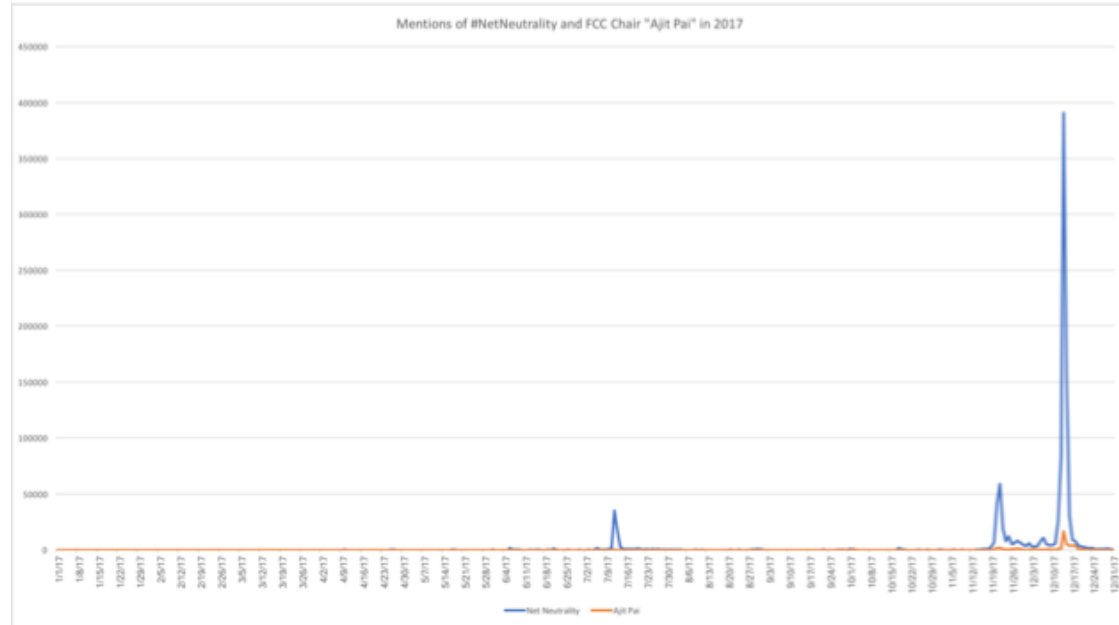


Blanks/gaps = interpolation (and implicit uncertainty)



Using the same data in different approaches

The same data will require different formatting & manipulation depending on your question



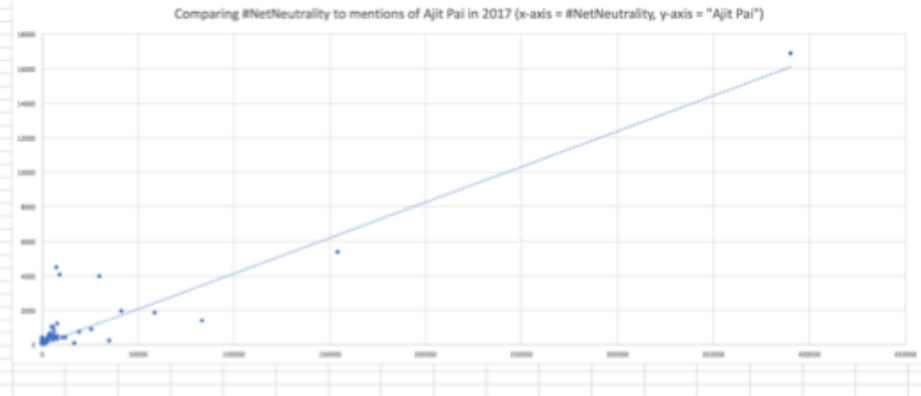
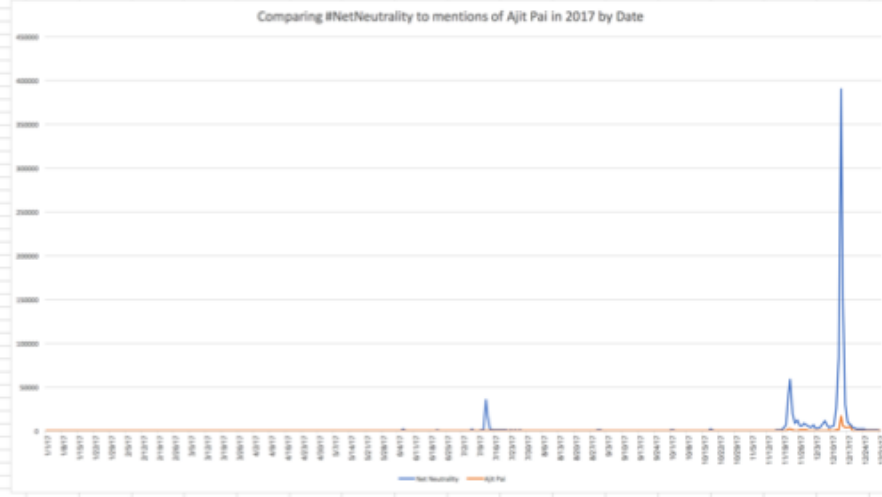
Hands on regression and scatterplot/trendline visualization

Open <http://tiny.cc/twitter-data>

Considerations:

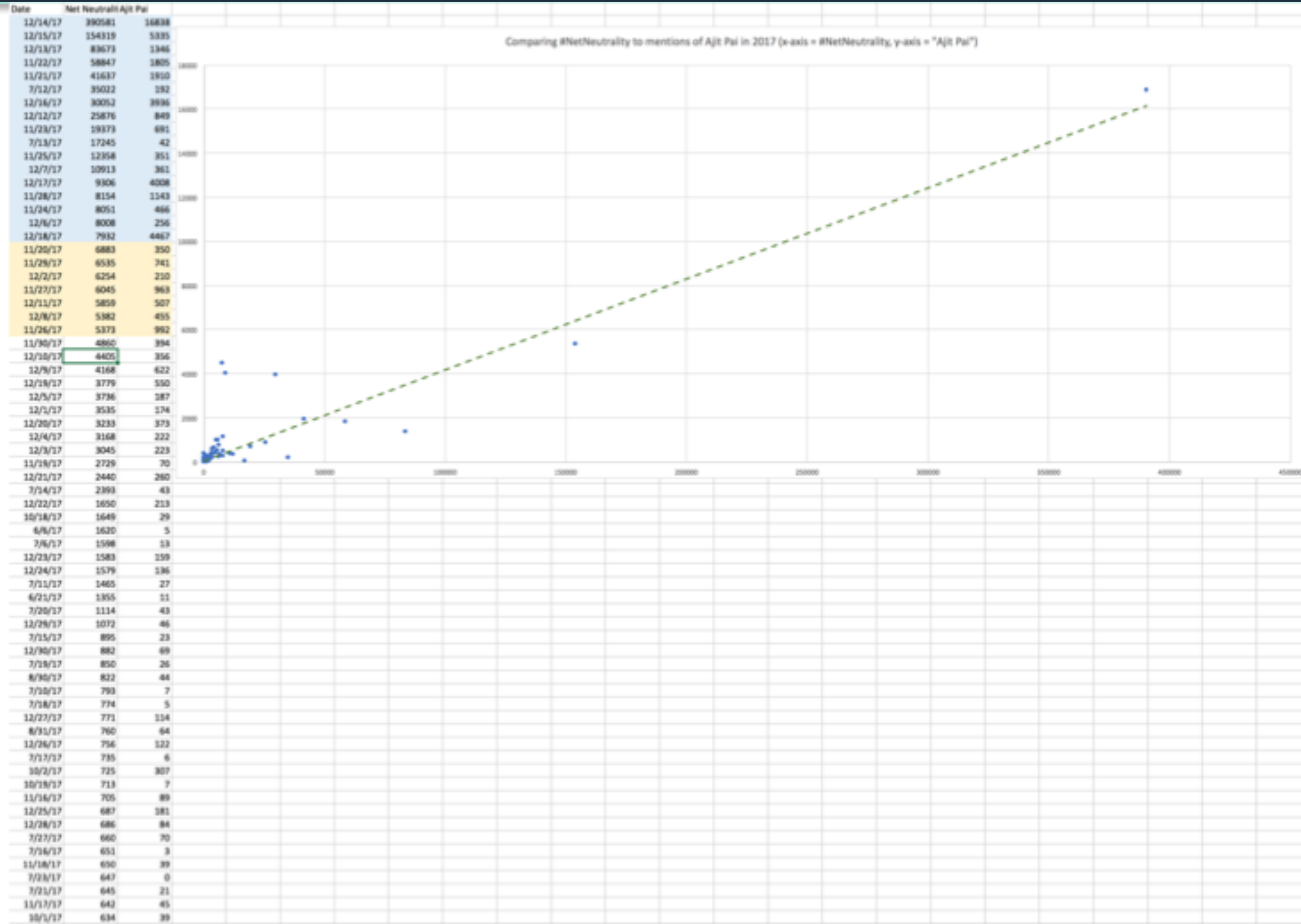
1. Which variable predicts (X) or is predicted (Y)?
2. Does your data meet (most) regression analysis assumptions?
 1. It's roughly linear (quadratic curves can be linear)
 2. Serious outliers have been removed
 3. Data is heteroscedastic (unevenly distributed; it will look conical in a scatterplot from an x/y 0/0 axis)

Date	Net Neutrality	Ajit Pai
1/1/17	0	0
1/2/17	3	0
1/3/17	4	0
1/4/17	10	0
1/5/17	4	0
1/6/17	4	0
1/7/17	0	0
1/8/17	2	0
1/9/17	0	0
1/10/17	0	0
1/11/17	3	0
1/12/17	5	0
1/13/17	4	0
1/14/17	2	0
1/15/17	1	0
1/16/17	9	0
1/17/17	0	2
1/18/17	7	1
1/19/17	3	0
1/20/17	0	26
1/21/17	5	8
1/22/17	1	4
1/23/17	22	167
1/24/17	34	96
1/25/17	27	17
1/26/17	24	4
1/27/17	10	7
1/28/17	5	3
1/29/17	2	2
1/30/17	4	5
1/31/17	32	15
2/1/17	24	8
2/2/17	6	1
2/3/17	8	6
2/4/17	8	8
2/5/17	5	1
2/6/17	5	9
2/7/17	115	3
2/8/17	56	1
2/9/17	12	1
2/10/17	1	0
2/11/17	8	5
2/12/17	11	3
2/13/17	5	2
2/14/17	11	1
2/15/17	15	2
2/16/17	4	0
2/17/17	3	0
2/18/17	7	0
2/19/17	4	7
2/20/17	14	3
2/21/17	8	8
2/22/17	0	1
2/23/17	7	1
2/24/17	16	2
2/25/17	6	0
2/26/17	6	1
2/27/17	28	3
2/28/17	27	15
3/1/17	12	0
3/2/17	10	2
3/3/17	12	0
3/4/17	5	0
3/5/17	5	0
3/6/17	5	1
3/7/17	3	8
3/8/17	17	5
3/9/17	6	5
3/10/17	7	3
3/11/17	7	0



Step 1: Get familiar with the data

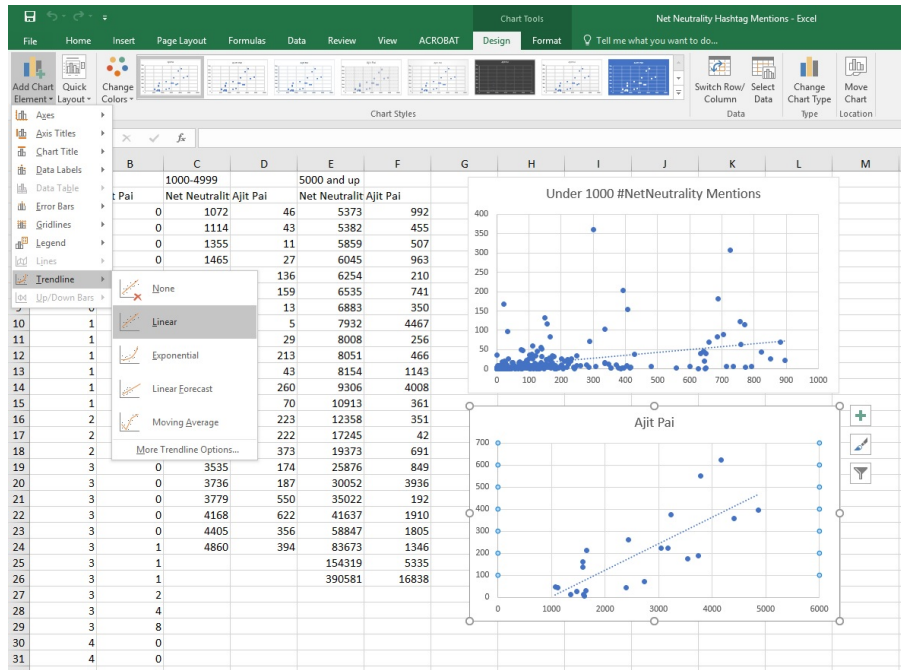
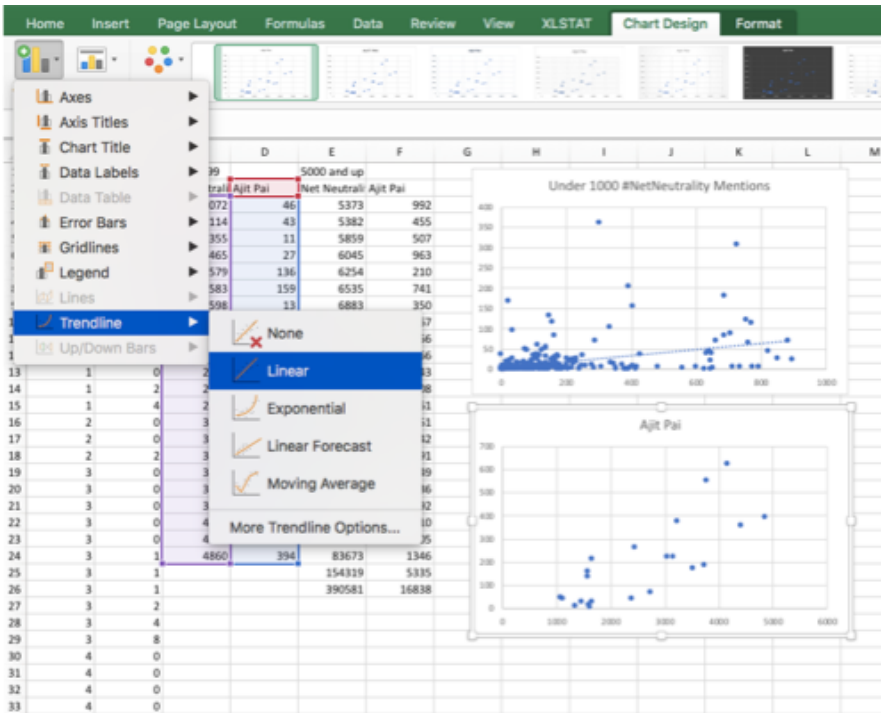
- Use scatterplots for data insight
- Assess linearity: Add your own trendline!



A hands-on break: Add your own trendline

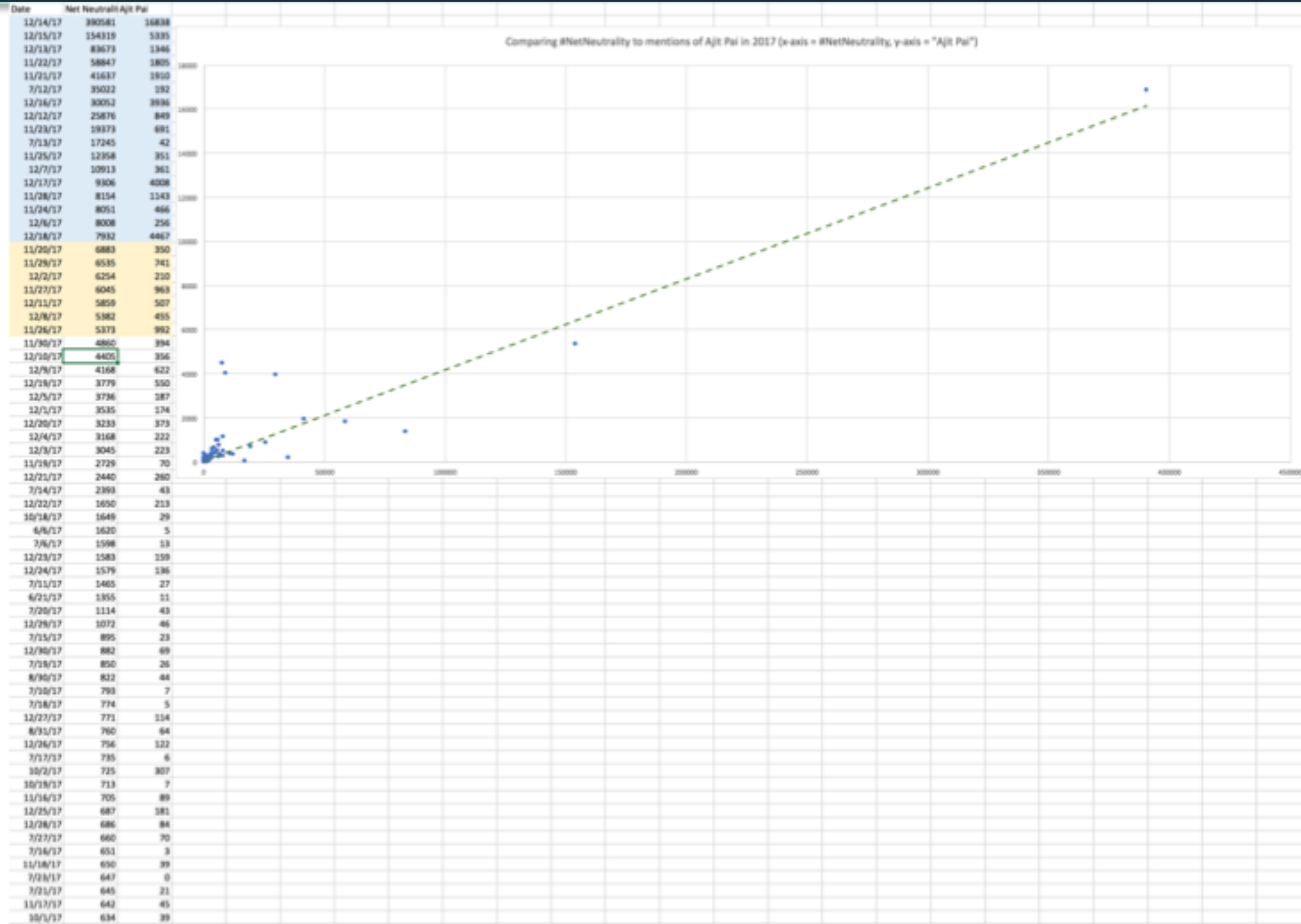
Mac: click on your chart, select Chart Design tab → Add Chart Element → Trendline

PC: click on your chart, click Design tab → Add Chart Element → Trendline

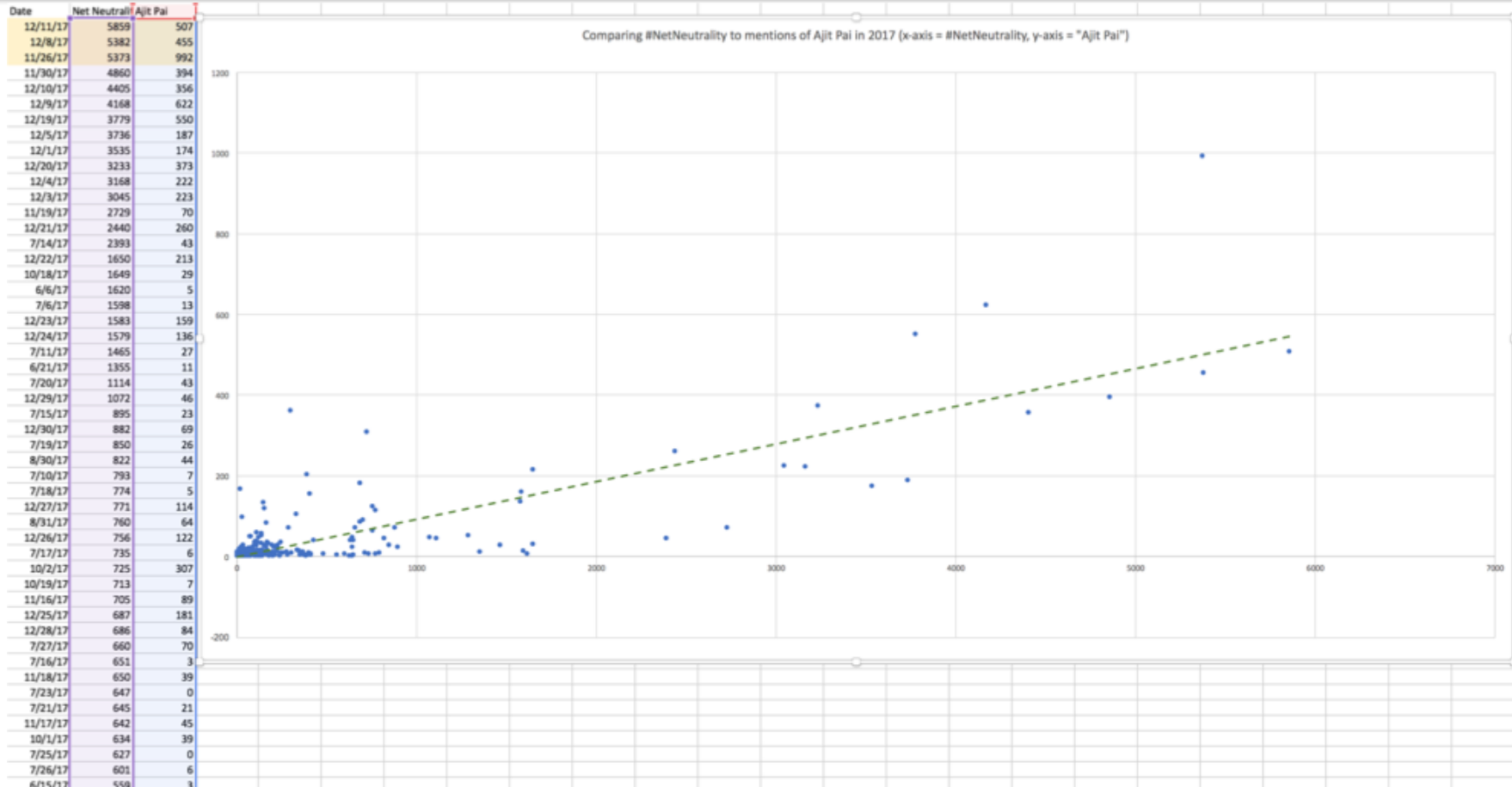


Step 1: Get familiar with the data

- Use scatterplots for data insight
- Assess linearity: Add your own trendline!
- Identify possible outliers and divisions in data



Step 2: Assess linearity and remove outliers



A hands-on break: Run your regression

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Under 1000		1000-4999		5000 and up								
2	Net Neutral	Ajit Pai	Net Neutral	Ajit Pai	Net Neutral								
3	0	0	1072	46	5373								
4	0	0	1114	43	5382								
5	0	0	1355	11	5859								
6	0	0	1465	27	6045								
7	0	1	1579	136	6254								
8	0	2	1583	159	6535								
9	0	36	1598	13	6883								
10	1	0	1620	5	7932								
11	1	0	1649	29	8008								
12	1	0	1650	213	8051								
13	1	0	2393	43	8154								
14	1	2	2440	260	9306								
15	1	4	2729	70	10913								
16	2	0	3045	223	12358								
17	2	0	3168	222	17245								
18	2	2	3233	373	19373								
19	3	0	3535	174	25876								
20	3	0	3736	187	30052								
21	3	0	3779	550	35022								
22	3	0	4168	622	41637								
23	3	0	4405	356	58847								
24	3	1	4860	394	83673								
25	3	1			154319								
26	3	1											
27	3	2											
28	3	4											
29	3	8											
30	4	0											
31	4	0											
32	4	0											

Regression

Input

Input Y Range:

Input X Range:

Labels Constant is Zero

Confidence Level: %

Output options

Output Range:

New Worksheet Ply:

New Workbook

Residuals

Residuals Residual Plots

Standardized Residuals Line Fit Plots

Normal Probability

Normal Probability Plots

OK Cancel

23R x 1C

Step 4: Understanding the regression

A	B	C
1000-4999 #NetNeutrality Mentions		
Regression Statistics		
Multiple R	0.787858886	
R Square	0.620721625	
Adjusted R Square	0.601757706	
Standard Error	736.0319651	
Observations	22	

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.391833644
R Square	0.153533604
Adjusted R Square	0.150863363
Standard Error	181.5151442
Observations	319

Significance F

3.77484E-13

	Coefficients	Standard Error	t Stat	P-value
Intercept	111.8135159	10.99784998	10.1668523	3.34748E-21
Ajit Pai	2.057363095	0.271321652	7.58274573	3.77484E-13

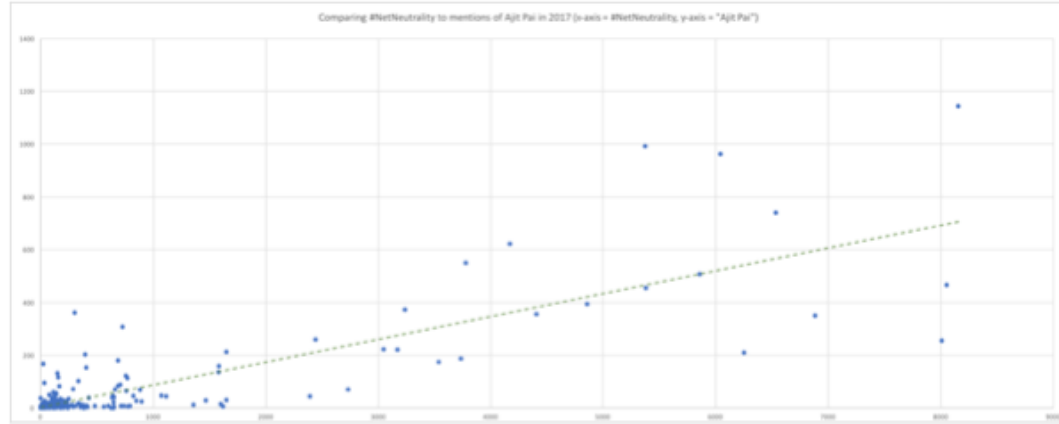
ANOVA

	df	SS	MS	F	Significance F
Regression	1	17732180.38	17732180.38	32.73171712	1.34199E-05
Residual	20	10834861.07	541743.0537		
Total	21	28567041.45			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1573.519675	232.2979883	6.773712018	1.37678E-06	1088.954562	2058.08479	1088.95456	2058.08479
Ajit Pai	5.187335697	0.906692367	5.721163965	1.34199E-05	3.296008562	7.07866283	3.29600856	7.07866283

Regression Analysis: Moving Forward

- The elephant in the room: prediction vs causation
- Visualizations encourage us to look for a relationship; statistical tests can confirm.
- Consider how you will:
 - Use this test to answer your research question
 - Split the dataset if necessary
 - Clean and transform your data to fit this test



Formulating Your Question:

- Is the dependent variable *influenced by* the independent variable?
- Does an increase in X predict an increase in Y?

Hands-on network analysis

A2482

AjitPaiFCC

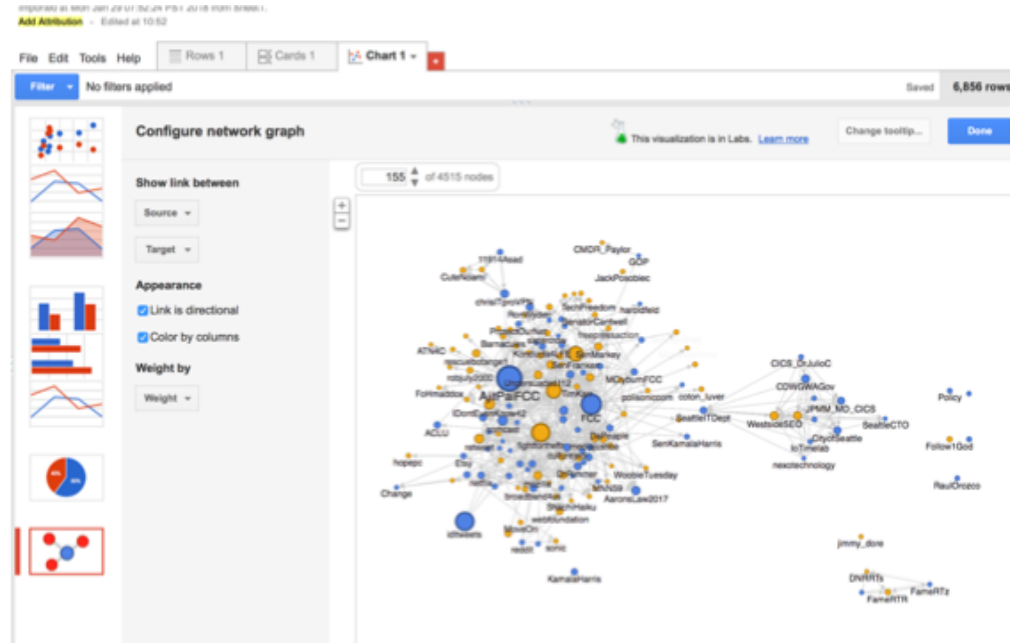
	A	B	C	D	E	F
		Source		Target	Weight	
1						
2	1summerstar1	1188	EntheosShin	18992	1	
3	1summerstar1	1188	retweet	50627	1	
4	61Randazxo	1335	SenatorCant	53939	1	
5	61Randazxo	1335	retweet	50627	1	
6	AjitPaiFCC	2482	DonnasueJ	16845	1	
7	AjitPaiFCC	2482	JeanZwier	28546	1	
8	AjitPaiFCC	2482	NatureGemir	43028	1	
9	AjitPaiFCC	2482	Taigitsune	58393	1	
10	AjitPaiFCC	2482	audiovoices	5505	1	
11	AjitPaiFCC	2482	fightfortheft	20369	33	
12	AjitPaiFCC	2482	freepress	21217	53	
13	AjitPaiFCC	2482	iammikegent	25699	2	
14	AjitPaiFCC	2482	ityiws	27180	1	
15	AjitPaiFCC	2482	janelle98	27880	1	
16	AjitPaiFCC	2482	jessmerica	29042	1	
17	AjitPaiFCC	2482	jimjames	29301	1	
18	AjitPaiFCC	2482	princeeditc			

1	___Daphne	1
2	___feah	2
3	___SlimGOODY	3
4	___Susan__	4
5	___tao	5
6	___Taty	6
7	___huma	7
2480	AjithVyaas	2480
2481	ajitpai	2481
2482	AjitPaiFCC	2482
2483	ajknight51	2483
2484	ajlfx	2484
2485	AjnaHorvath	2485
2486	AJNerds wag	2486
2487	ajohnsoncook	2487
2488	ajolie11	2488
2489	AJoos	2489
	uffo	2490

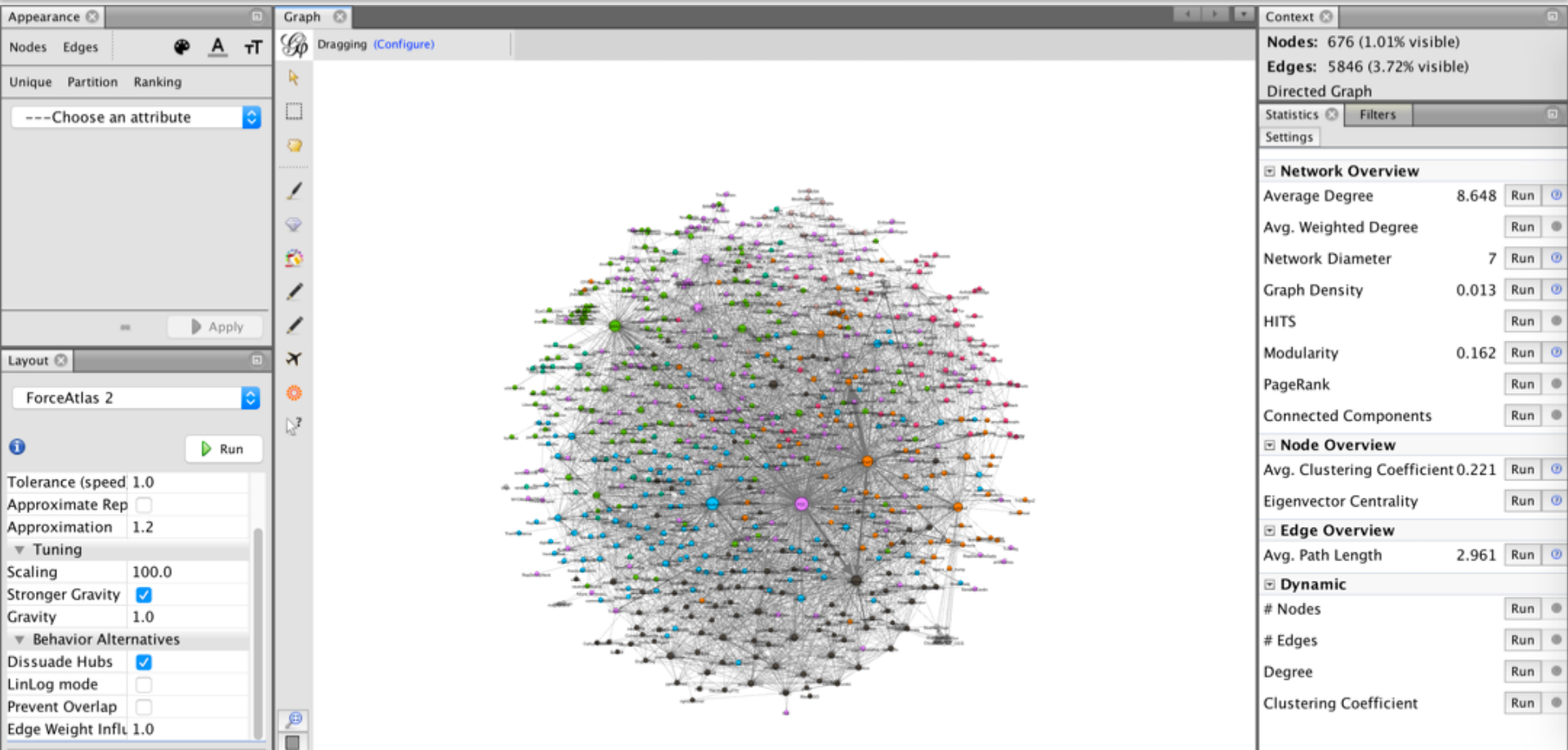
Open <http://tiny.cc/network-data>

Using Google spreadsheets with Google Fusion Tables

- Requirements
 - Source and Target fields in a Google Drive spreadsheet
 - Non iu.edu account
- Strengths
 - Interactive gravity
 - Fast and easy
- Limits
 - 1000 nodes, 5000-ish edges (?)
 - Only one layout type (Yifan Hu)
 - No stats



Statistics underpin network visualizations



The screenshot displays a network visualization software interface. The central area shows a large, dense network graph with numerous nodes and edges. The nodes are colored in various colors (red, green, blue, purple, orange, black) and are connected by a complex web of edges. The interface is divided into several panels:

- Appearance:** Contains tabs for Nodes and Edges. A dropdown menu shows "--Choose an attribute". An "Apply" button is visible.
- Layout:** Shows the selected layout as "ForceAtlas 2". A "Run" button is present.
- Settings:** Includes sections for "Tuning" and "Behavior Alternatives".
 - Tuning:** Scaling: 100.0; Stronger Gravity: ; Gravity: 1.0.
 - Behavior Alternatives:** Dissuade Hubs: ; LinLog mode: ; Prevent Overlap: ; Edge Weight Infl: 1.0.
- Context:** Provides summary statistics:
 - Nodes:** 676 (1.01% visible)
 - Edges:** 5846 (3.72% visible)
 - Directed Graph**
 - Statistics:** Includes a "Filters" tab.
 - Network Overview:**
 - Average Degree: 8.648 (Run)
 - Avg. Weighted Degree: (Run)
 - Network Diameter: 7 (Run)
 - Graph Density: 0.013 (Run)
 - HITS: (Run)
 - Modularity: 0.162 (Run)
 - PageRank: (Run)
 - Connected Components: (Run)
 - Node Overview:**
 - Avg. Clustering Coefficient: 0.221 (Run)
 - Eigenvector Centrality: (Run)
 - Edge Overview:**
 - Avg. Path Length: 2.961 (Run)
 - Dynamic:**
 - # Nodes: (Run)
 - # Edges: (Run)
 - Degree: (Run)
 - Clustering Coefficient: (Run)

Data format will vary by application

Time series

Date	Net Neutrality	Ajit Pai
11/21/17	41637	1910
11/22/17	58847	1805
11/23/17	19373	691
11/24/17	8051	466
11/25/17	12358	351
11/26/17	5373	992
11/27/17	6045	963
11/28/17	8154	1143
11/29/17	6535	741
11/30/17	4860	394
12/1/17	3535	174
12/2/17	6254	210
12/3/17	3045	223
12/4/17	3168	222
12/5/17	3736	187
12/6/17	8008	256
12/7/17	10913	361
12/8/17	5382	455

Regression

Net Neutrality	Ajit Pai
41637	1910
58847	1805
19373	691
8051	466
12358	351
5373	992
6045	963
8154	1143
6535	741
4860	394
3535	174
6254	210
3045	223
3168	222
3736	187
8008	256
10913	361
5382	455

Network analysis

Target	Source
1summerstar1	EntheosShines
61Randazxo	SenatorCantwel
AjitPaiFCC	DonnasueJ
AjitPaiFCC	JeanZwier
AjitPaiFCC	NatureGemini
AjitPaiFCC	Taigitsune
AjitPaiFCC	audiovoices
AjitPaiFCC	fightfortheft
AjitPaiFCC	fightfortheft
AjitPaiFCC	freepress
AjitPaiFCC	iammikegentile
AjitPaiFCC	iammikegentile
AjitPaiFCC	ityiws
AjitPaiFCC	janelle98
AjitPaiFCC	jessmerica
AjitPaiFCC	jimjames
AjitPaiFCC	princeeditor

Geographic Analysis

lat	long	count
-0.02	37.9	8
-1.78	-78.09	3
-12.04	-77.02	5
-15.77	-47.92	2
-24.91	133.39	8
-28.48	24.67	4
-33.46	-70.64	1
-37.81	144.96	1
-43.58	170.36	7
1.36	103.82	8
10.49	-66.89	3
17.43	78.5	1
18.22	-66.42	10
18.4	-66.1	1
18.42	-66.06	2
19.43	-99.13	5
21.02	105.8	2
21.78	82.79	20
24.28	55.11	1
27.01	-80.45	5

IU Resources

Dr. Emily Meanwell

Director, Social Science Research Commons

emeanwel@indiana.edu

David Kloster

Programmer/Analyst, Cyberinfrastructure for Digital
Humanities

klosteda@iu.edu

Further Resources

Data & Methods Consults

- IDAH Consultation Hours
 - Tues 10a-12p; Wed 2-4p; Fri 10a-12p
 - Email idah@indiana.edu
- Center for Survey Research
 - Stacey Giroux: sagiroux@indiana.edu
- Social Science Research Commons
 - <http://ssrc.indiana.edu>
- Online Tutorials
 - <https://statistics.laerd.com>

Upcoming IDAH Events:

“Virtual Realities: Making the Humanities in a Digital World”

A talk by Dr. William “Bro” Adams, former head of the National Endowment for the Humanities

Tuesday, February 20, 5:30pm
Woodburn Hall 100

Computational Humanities Reading Group February’s Topic: “Against Data Analysis”

Friday, February 16, 1pm
IDAH Conference Room E171