

UNIT – III

Case Study – I

Email, Thread Networks and Twitter

By

Ranjith Kumar M

Assistant Professor, CSE

SRITW

Content

- **Email**
 - **The lifeblood of Modern Communication**
- **Thread**
 - **Mapping Message Boards and Email Lists**
- **Twitter**
 - **Conservation**
 - **Entertainment and Information**

Email

- Electronic mail, or email, is an electronic message transmitted over a communications network, typically as a text file with optional attachments. Email is older than the Internet itself.
- Email services became essentially free from an end-user perspective with popular web services such as Hotmail, Yahoo mail, and Gmail.
- Some important technical characteristics make email particularly powerful:
 - *Flexible form.* Email can be a simple plain-text message, a richly formatted newsletter, or even an interactive survey.
 - *Asynchronous.* The asynchronous nature of email allows people to send and receive messages on their own time, without interrupting others.

Email

- Some important technical characteristics make email particularly powerful:
 - *Broadcast*. Emails can be sent to any number of people simultaneously. Ad hoc groups can be created on the fly by sending to multiple email addresses and using the common Reply to All feature.
 - *Push technology*. Email is considered a push technology; the sender decides what shows up in the receiver's inbox without any action on the receiver's end. This is great when trying to get someone's attention, but is also the reason so much unwanted email spam gets sent around.
 - *Threaded conversation*. Email messages are often organized into a threaded pattern consisting of messages, replies to messages, replies to replies, and so forth.

Email

- Email Network
 - In a standard email network, vertices represent email addresses or corresponding people.
 - Edges or ties are created when a message is sent from one email address to another.
 - Edges are directed because messages are transferred from a sender to a receiver.
 - These ties are weighted by the number of messages sent between two individuals.

Email

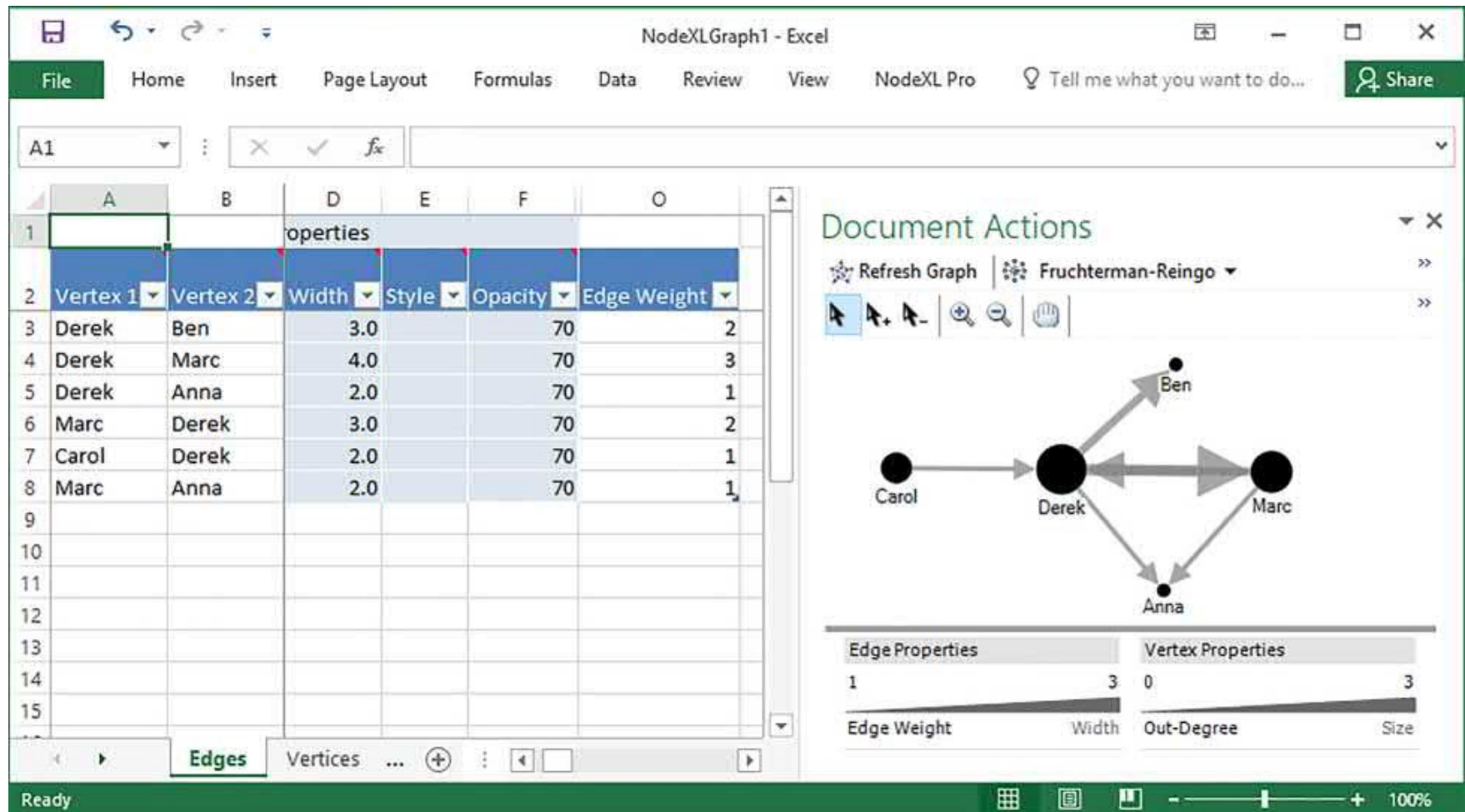
- Email Network

| From | To | Cc | Subject |
|-------|-----------|------|---------------------|
| Derek | Ben | | HCIL Brownbag |
| Derek | Marc, Ben | | Travel Plans |
| Derek | Marc | Anna | Registration |
| Marc | Derek | | Re: Travel Plans |
| Carol | Derek | | Tuesday meeting |
| Marc | Derek | Anna | Re: Registration |
| Marc | Derek | | Next Steps |

This email network edge list contains seven messages from Derek's personal email collection that includes ten unique edges (including both To and Cc) and five vertices.

Email

- Email Network



Email

- **What questions can be answered by analyzing email networks?**
- Three main types of email collections (personal, organizational, and community), each of which may be analyzed by a current participant or an outside observer.
 - Personal email collections include messages sent or received by an individual.
 - Organizational email collections include messages sent and received by members of an organization. More generally, they are the aggregate of several individuals' personal email collections.
 - Community email collections include messages sent to an email list address that get forwarded to a group of subscribed members.

Email

- **What questions can be answered by analyzing email networks?**

| | Personal | Organizational | Community |
|---------------------|---------------------------------------------------|---------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| Current Participant | Region A: Analyzing your own email | Region B: Analyzing your organization's email | Region C: Analyzing ongoing conversations in a community email list in which you participate |
| Outside Observer | Region D: Analyzing another person's email | Region E: Analyzing another organization's email | Region F: Analyzing a community email list archive in which you do not participate |

Email

- **Personal email network questions**
- Several questions can be asked about personal email network datasets:
 - *Individuals.* Who are important individuals within the network?
 - For example, who are boundary spanners who link across clusters of contacts?
 - Who is contacted most often?
 - Who are the most active discussants of a particular topic?
 - Who are unwanted or troublesome correspondents?
 - Who are unresponsive recipients?

Email

- **Personal email network questions**
- Several questions can be asked about personal email network datasets:
 - **Groups.** *What natural subgroups exist?*
 - *What collaborative activities are individuals engaged in?*
 - *What are the relationships between subgroups?*
 - **Temporal comparisons.** *How have relationships changed over time?*
 - *How did an event (e.g., move to a new location) affect the network?*
 - *What inactive groups exist with whom I may benefit from reestablishing contact?*
 - *What projects or people have I neglected?*

Email

- **Personal email network questions**
- Several questions can be asked about personal email network datasets:
 - ***Structural patterns.*** Are there common social roles that occur among contacts (e.g., informant, decision maker, boundary spanner)?
 - Are there types of subgroups that occur (e.g., cliques, fans)?

Email

- **Organizational email network questions**
- Several different questions can be asked about organizational email network datasets:
 - **Individuals.** *Who are the important individuals* within an organization?
 - For example, who are the boundary spanners who link across organizational silos?
 - Who are the influencers or topical experts?
 - Who is not well connected and could benefit from more social ties?
 - Who would be a good replacement for an individual?
 - Who fills a unique niche?
 - Who was in-the-know about an important decision?

Email

- **Organizational email network questions**
- Several different questions can be asked about organizational email network datasets:
 - **Groups.** How do email-based groupings differ from organizational structures?
 - How is the “org-chart” different from the chart of the flow of email?
 - How are groups interconnected?
 - Which groups should be better connected? Is there a core competency that is not discussed in a particular branch or office?

Email

- **Organizational email network questions**
- Several different questions can be asked about organizational email network datasets:
 - ***Temporal comparisons.*** How does information flow through the organization?
 - How do connections among individuals and subgroups evolve over time?
 - How are social relations affected by a major event such as a merger or opening of a new office?
 - ***Structural patterns.*** What network properties are related to success? Can we identify up-and-coming stars or unique social roles based on their network structure? How is information on a particular topic distributed throughout the organization?

Email

- **Working with email data**
- The email header includes the From, To, Cc, Bcc, Date, and Subject fields.
- The email body includes the message content and any attachments.
- Email is transmitted through the Internet via the Simple Mail Transfer Protocol (SMTP).
- Email uses the Multipurpose Internet Mail Extensions (MIME) format to allow character sets other than ASCII and non-text attachments to be included and transported via email.

Email

- **Working with email data**
- Email client applications such as Microsoft Outlook or Apple Mail retrieve or cache messages from a mail server using Post Office Protocol (POP) or Internet Message Access Protocol (IMAP).
- Corporate email is typically retrieved through proprietary protocols specific to Microsoft Exchange Servers or competitors.
- Email messages are stored in a variety of formats for different email clients. Some email clients store each message as a separate file; others save them in a database format.

Email

- **Working with email data**
- Some common formats include .eml (Microsoft Outlook, Mozilla Thunderbird), .emlx (Apple Mail), .msg and .pst (Microsoft Outlook or Microsoft Exchange), and .mbox (Mozilla Thunderbird, Gmail backup files, and many email list archives).

Email

- **Working with email data**
- Working with email poses technical challenges that often require preprocessing data to create useful results.
- The large potential size of email networks can be problematic and may require specialized programs to manage large data volumes.
- In practice, email will likely need to be filtered before analysis to reduce the dataset based on time ranges, people, and topics of interest.
- Another major challenge is the use of multiple email addresses for the same individual.

Email

- **Working with email data**
- In most cases, analysts are interested in social relationships between individuals, not the relationships between email accounts.
- The problem of combining different aliases (email addresses) for the same entity (person) is called “entity resolution,” “identity resolution,” “deduplication,” or “record linkage.”
- A range of tools provide deduplication services such as Marketo or the open source Python library Dedupe.

Email

- **Working with email data**
- Another set of tools extracts entities (e.g., names or places mentioned in email messages) which can be used to create networks that consider personal names or places mentioned in email texts rather than the sender and receiver of a message.
- Searching for tools that perform “named-entity recognition,” “entity identification,” “entity extraction” and “entity chunking” reveals tools such as the spaCy Python library, Stanford NER, and commercial APIs such as Lexalytics, TextRazor, ParallelDots, and Aylien.

Email

- **Preparing email**
- The easiest way to transform email messages into network relationships (i.e., an edge list) is to use NodeXL's Import from Email Network feature.
- This feature relies on the Windows built-in indexing functionality on recent versions of Windows (e.g., Windows 10). By default, email files in certain formats will be indexed by Windows.
- You can view and change which filetypes are indexed and check indexing progress in the Indexing Options dialog accessible via the Control Panel.

Email

- **Importing email network into NodeXL**
- Once Windows has indexed the email you want to analyze, you are ready to import the data directly into NodeXL.
- Select the From Email Network option from the Import drop-down on the NodeXL ribbon to open the importer
- The enormous size of many email collections often requires filtering out messages before analysis.
- There are several ways of filtering:
 - *Filter based on time.*
 - *Filter based on sender and receiver(s)*
 - *Filter based on content.*
 - *Filter based on folders and labels.*
 - *Filter based on a combination of features.*

Email

- **Importing email network into NodeXL**

Import from Email Network ✕

☐ Analyze all emails [How email is analyzed and imported](#)

☒ Analyze filtered emails only

Filters

☐ Includes these email addresses on the From, To, Cc, or Bcc lines
[About email addresses](#)

| | Email Address | From | To | Cc | Bcc |
|---|---------------|--------------------------|--------------------------|--------------------------|--------------------------|
| * | | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

Subject
☐ Includes text:

Message body
☒ Includes text:

Date range
☒ Sent on or after:
☒ Sent on or before:

Size range (bytes)
☐ Minimum:
☐ Maximum:

☐ Attachments
☒ Has attachments
☐ Doesn't have attachments
☐ From first email address and has attachments

Folder
☐ In folder:
[Sample folders](#)

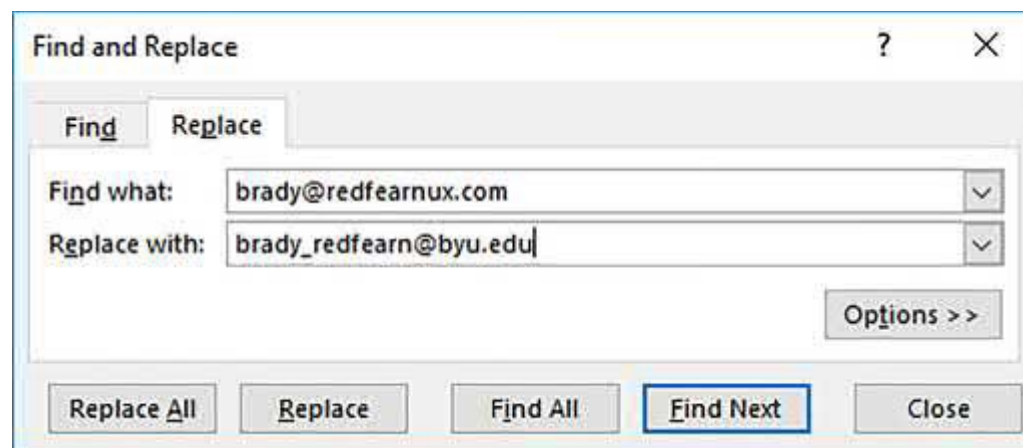
☐ Cc
☒ Has Cc
☐ Doesn't have Cc

☐ Bcc
☒ Has Bcc
☐ Doesn't have Bcc

☒ Use Cc line when calculating edge weights
☐ Use Bcc line when calculating edge weights

Email

- **Cleaning email data in NodeXL**
- After importing email data into NodeXL, you will likely need to clean it to remove duplicate email addresses for the same individuals, as well as self-referring loops created when people reply to their own messages.
 - Remove duplicate email addresses from the same individuals
 - Count and merge duplicate edges



Email

- **Analyzing personal email network**
- *Import data into NodeXL*
- *Clean data*
- *Filter data*
- *Compute graph metrics and add new Columns*
- *Visualize the email social network*
- *Understand social network visualizations and metrics data*

Email

- **Creating an expertise network email graph**
- *Import email social network data into NodeXL*
- *Clean data*
- *Compute graph metrics and add new Columns*
- *Filter data*
- *Visualize network*
- *Understanding the network visualization and data*

Thread Network

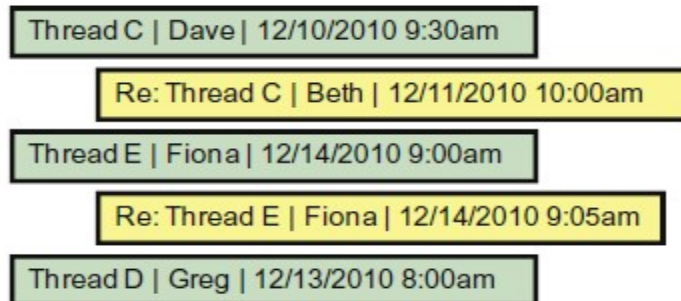
- **Mapping Message Boards and Email Lists**
- Threaded conversation is a commonly used design theme that enables online discussion between multiple participants using the ubiquitous post-reply-reply structure.
- The key properties of threaded conversation were enumerated in Resnick et al., and are listed here with some modification:
 - Topics
 - Threads
 - Single Authored
 - Permanence
 - Thread Navigation

Thread Network

TOPIC 1: Social Media



TOPIC 2: NodeXL



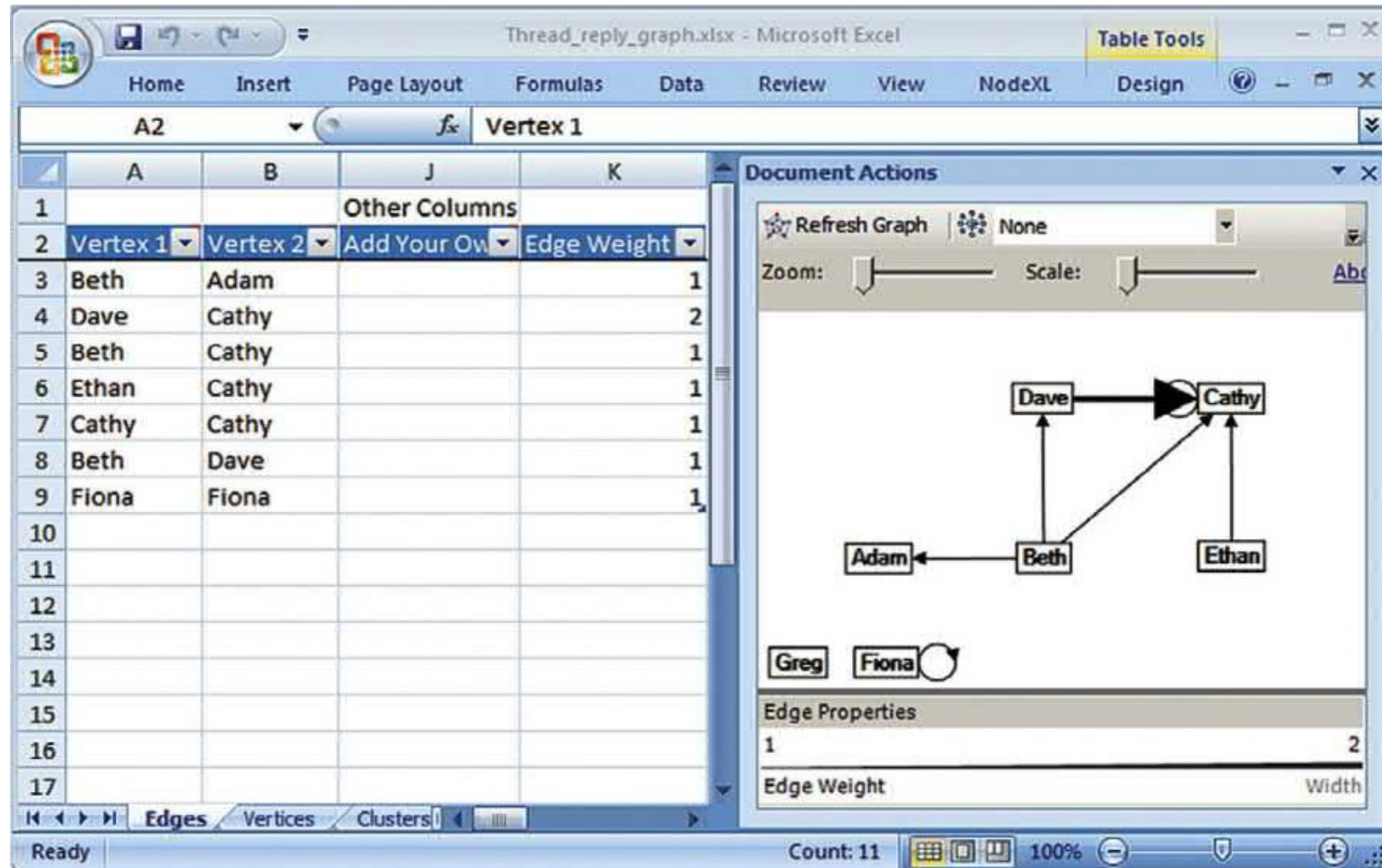
Thread Network

- **Mapping Message Boards and Email Lists (Questions)**
- **Individuals** : Who are important individuals within the community?
- **Groups** : Who makes up the core members of the community?
- **Temporal comparisons** : How have participation patterns and overall structural characteristics of the community changed over time?
- **Structural Pattern** : What network properties are related to community sustainability?

Thread Network

- **Threaded Conversation network**
- The network most commonly used to analyze threaded conversations is the Reply network.
- Each time someone replies to another person's message

Thread Network



Thread Network

- **Identifying important people and social roles in the CSS-D Q&A reply network**
- There are a host of email lists, forums, and Q&A websites such as Stack overflow and Quora where people post technical questions and volunteers provide answers.
- Many companies host these Q&A discussions to learn about problems with existing products, resolve customer concerns, generate new ideas on future improvements, and build a loyal customer community.

Thread Network

- **Identifying important people and social roles in the CSS-D Q&A reply network**
- *Identify high-value contributors of different types.*
- *Determine if your community has the right mix of people.*
- *Recognize changes and vulnerabilities in the social space.*

Twitter

- **Information flows, influencers, and organic communities**
- In social media, users form social networks by making relationships with other users.
- On Twitter, users make relationships when they follow other users, mention users in a tweet, reply to a tweet, or “retweet” another tweet.
- Once you follow another user by subscribing to their posted content, this content will then appear on your Twitter feed.

Twitter

- Twitter conversations span a wide range of topic and issues.
- In order to collect topic-specific Twitter network data, you must first define your topic.
- State your topic first (e.g., Soft drinks and obesity, or the name of a political party, celebrity, or product).
- Next, explore related keywords and hashtags using the built-in Twitter Search and third party apps such as <https://hashtagify.me> to identify trends, popularity, and related terms.
- In order to communicate your topic to Twitter, you will use a Boolean search query that combines these terms into a specific machine-readable format.

Twitter

- A Boolean search is a query technique that utilizes Boolean Logic to connect individual keywords or phrases within a single query.
- The term “Boolean” refers to a system of logic developed by the mathematician and early computer pioneer, George Boole. Boolean searching includes three key Boolean operators: AND, OR, and NOT.
- An **AND** operator *narrows* your search. Between two keywords it results in a search for posts containing both of the words.
- Boolean search “Cats AND Dogs” will retrieve all posts that contain **both** words.
- A lack of an operator between words is interpreted as an AND operator within NodeXL. For example: “Cats Dogs” is treated the same as “Cats AND Dogs”.

Twitter

- An **OR** operator *expands* your search. An OR operator will return any posts that contain at least one of the search terms.
- Boolean search “Cats OR Dogs” will retrieve posts that
- contain either word. Example returned post: “love all dogs!”
- A **NOT** operator *excludes* posts containing the keyword. Using the NOT operator will exclude any posts containing the keyword *following* the operator.
- Boolean search “Cats NOT Dogs” will retrieve all posts that have the word “Cats” in them, unless it also has the word “Dogs.”
- Twitter also recognizes the minus sign (–) as a NOT operator
- (E.g.” Cats -Dogs”) within Boolean searches. When collecting data using NodeXL you should use the “-” operator rather than the word NOT.

Twitter

- Two other indicators that are helpful in constructing a Boolean search are parentheses and quotation marks.
- **Quotation** marks requires words to be searched as a phrase, in the exact order you type them. For example, searching for posts about the movie *Love Actually*, using the quotation marks will search for the phrase exactly as it appears (i.e., “Love Actually”).
- This post will be *included*: “Hugh Grant is a great fit for his role in Love Actually.” This post will not be *included*: “Actually, falling in love is not as simple as you may think.”
- **Parentheses** require the terms and operations that occur inside them to be searched first. Sometimes called *nesting*, parentheses add a level of organization for your Boolean search, allowing you to formulate complex search strings.

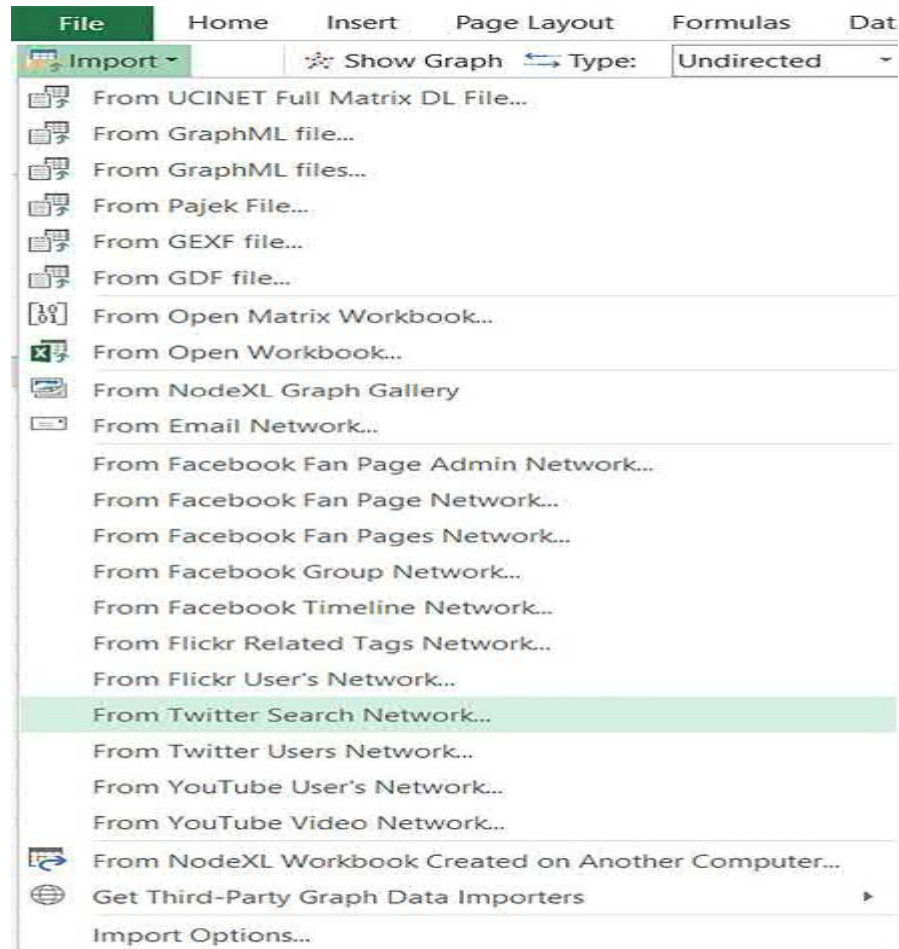
Twitter

- Consider the following Boolean search:
“(Cats OR Kittens) AND (Dogs OR Puppies).”
- The search will first be performed within the parentheses, and only then between them. Results would include any post that contains a combination of the two nested boolean queries.

Twitter

- **Twitter Data Collection**
- Open NodeXL Pro and click on the NodeXL Pro tab along the top menu ribbon in Excel.
- Click Import in the Data section of the NodeXL ribbon. In the drop-down menu, choose From Twitter Search Network
- You can use the NodeXL Twitter Search Network importer dialog to extract networks for your search string.
- Several options are available to refine data requests using this importer.
- The NodeXL data collector starts by performing a query against the Twitter Search service at <http://search.twitter.com>.

Twitter



Twitter

- **The raw data layout**
- The raw Twitter data you have just downloaded will populate the Edges and Vertices worksheets.
- In this Twitter analysis, the Edges worksheet contains a row for each edge which represents a connection event between two people who tweeted within the collected dataset.
- Edges represent the various kinds of relationships that can be created through Twitter.
- NodeXL constructs four different types of Twitter edges: Mentions, Replies to, Retweet, and Tweet. A Mentions edge is created when one user creates a tweet that contains the name of another user, which is NOT at the beginning of the tweet

Twitter

Import from Twitter Search Network

[This might take a long time: Twitter rate limiting](#)

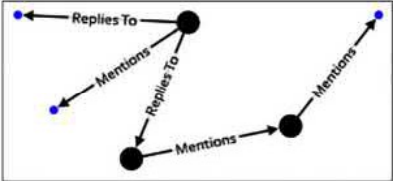
Search for tweets that match this query:

[How to use advanced search operators](#)

What to import

☒ Basic network
Show who was replied to or mentioned in recent tweets
[More about this option](#)

☐ Basic network plus friends (very slow!)
Add some of the users' friends
[More about this option](#)



Your Twitter account

☐ I have a Twitter account, but I have not yet authorized NodeXL to use my account to import Twitter networks. Take me to Twitter's authorization Web page.

☒ I have a Twitter account, and I have authorized NodeXL to use my account to import Twitter networks.

Limit to tweets

☒ Limit friends and followers to per user

☒ Expand URLs in tweets (slower)

☐ Extended analysis: perform a second pass on the collected Tweets to ensure that all Retweets are collected and all RetweetedIDs are correct. (Slow!)

OK

Cancel

Twitter

- Vertex 1 and Vertex 2 represents the two users connected by a tweet. Vertex 1 is the author of the tweet and Vertex 2 is the target of the tweet by virtue of being mentioned, replied-to, retweeted, or tweeted (in which case the same username shows up in both columns).
- Note that Twitter networks are Directed networks.
 - The Visual Properties and Labels columns are not part of the raw data, and therefore empty at the start of the analysis. You will learn more about these values after they are populated.
 - The type of connection, Mentions, Replies to, Retweet, and Tweet, is identified in the Relationship column.
 - Relationship Date refers to the date and time when the tweet was posted in Coordinated Universal Time (UTC), also known as Greenwich Mean Time (GMT-0).

Twitter

- Note that Twitter networks are Directed networks.
 - The Tweet column includes the text of the tweet itself.
 - The URLs column lists the full expanded hyperlinks of any shortened links mentioned in the tweet text.
- This is created when the expand URLs in Tweets option is selected.
- The Hashtags in Tweets column extracts the hashtags from the tweet for further analysis.
- The Vertex worksheet provides information about individual users in the network. Each row represents a

Twitter

Posting users

Relationship target

Relationship type

| | | | | | | | | | | | | | | | | | | | | | | |
|-----------------------------------------------------------------------------------------------------------------------------------------|--------------|-----------------|--|--|--|--|--|--|--|--|--|--|---------------------------|------------|-----------------|--|--------|--|----------------|--|-----------------|--|
| N4950 | | | | | | | | | | | | | Replies to | | | | | | | | | |
| Visual Properties | | | | | | | | | | | | | Labels | | Other Columns | | | | | | | |
| | | | | | | | | | | | | | Add Your Own Columns Here | | Relationship | | | | | | | |
| | | | | | | | | | | | | | Date (UTC) | | | | | | | | | |
| 4896 | truthserum | truthserumusa | | | | | | | | | | | | Tweet | 3/29/2012 19:23 | | | | | | | |
| 4897 | truthserum | truthserumusa | | | | | | | | | | | | Tweet | 3/29/2012 19:29 | | | | | | | |
| 4922 | kazydw | preciousliberty | | | | | | | | | | | | Mentions | 3/29/2012 19:06 | | | | | | | |
| 4923 | preciouslib | preciousliberty | | | | | | | | | | | | Tweet | 3/29/2012 19:05 | | | | | | | |
| 4924 | preciouslib | preciousliberty | | | | | | | | | | | | Tweet | 3/29/2012 19:21 | | | | | | | |
| 4925 | preciouslib | preciousliberty | | | | | | | | | | | | Tweet | 3/29/2012 19:40 | | | | | | | |
| 4926 | truevineflo | preciousliberty | | | | | | | | | | | | Mentions | 3/29/2012 19:29 | | | | | | | |
| 4940 | penguinpoi | rbpundit | | | | | | | | | | | | Mentions | 3/29/2012 19:38 | | | | | | | |
| 4941 | penguinpoi | melissatweets | | | | | | | | | | | | Mentions | 3/29/2012 19:38 | | | | | | | |
| 4949 | wes_wheel | thegrudgeratort | | | | | | | | | | | | Mentions | 3/29/2012 19:37 | | | | | | | |
| 4950 | wes_wheel | thegrudgeratort | | | | | | | | | | | | Replies to | 3/29/2012 19:37 | | | | | | | |
| 4951 | thegrudger | wes_wheeler | | | | | | | | | | | | Mentions | 3/29/2012 19:37 | | | | | | | |
| 4965 | olgold96 | coondawg68 | | | | | | | | | | | | Mentions | 3/29/2012 19:30 | | | | | | | |
| 4974 | foolishrepc | qc16 | | | | | | | | | | | | Mentions | 3/29/2012 19:40 | | | | | | | |
| 4988 | thegrudger | thegrudgeratort | | | | | | | | | | | | Tweet | 3/29/2012 19:09 | | | | | | | |
| 4989 | thegrudger | thegrudgeratort | | | | | | | | | | | | Tweet | 3/29/2012 19:41 | | | | | | | |
| 4990 | thegrudger | thegrudgeratort | | | | | | | | | | | | Tweet | 3/29/2012 19:41 | | | | | | | |
| 4991 | revenant01 | thegrudgeratort | | | | | | | | | | | | Mentions | 3/29/2012 19:41 | | | | | | | |
| 5009 | politicaltod | politicalodd1 | | | | | | | | | | | | Tweet | 3/29/2012 19:14 | | | | | | | |
| 5017 | truevineflo | truevineflorida | | | | | | | | | | | | Tweet | 3/29/2012 19:45 | | | | | | | |
| 5024 | sterlingvotl | sterlingvoth | | | | | | | | | | | | Tweet | 3/29/2012 19:06 | | | | | | | |
| 5036 | occupypolit | left4liberty | | | | | | | | | | | | Mentions | 3/29/2012 19:06 | | | | | | | |
| 5037 | occupypolit | occupypolitics | | | | | | | | | | | | Tweet | 3/29/2012 19:15 | | | | | | | |
| 5038 | occupypolit | torianbrown | | | | | | | | | | | | Mentions | 3/29/2012 19:18 | | | | | | | |
| 5039 | occupypolit | torianbrown | | | | | | | | | | | | Mentions | 3/29/2012 19:44 | | | | | | | |
| Edges | | | | | | | | | | | | | | | Vertices | | Groups | | Group Vertices | | Overall Metrics | |
| @preciousliberty: Arianna Huffington: Tired of #Obama rhetoric that never delivers http://t.co/k8QfPV2B #citizensunited #rs #p2 #edshow | | | | | | | | | | | | | | | | | | | | | | |

Mentions

Replies to

Retweet

Tweet – placeholder for non-relational tweets (creates a self-loop)

Relationship types

Mentions
Replies to
Retweet

Tweet – placeholder
for non-relational
tweets (creates a
self-loop)

Twitter

| Relationship | | URL | | #s | |
|--------------|--------------|-----------------|----------------------------------------------------------------------------------------------|----------|------------------|
| Vertex 1 | Vertex 2 | Date (UTC) | Tweet | Tweet | Tweet Date (UTC) |
| truthserum | truthserum | 3/29/2012 19:23 | FACT: Domestic Oil Production Is At Eight-Year High: http://t.co/http://t.co#election2012 | | 3/29/2012 19:23 |
| truthserum | truthserum | 3/29/2012 19:29 | Smear campaign on climate scientists http://t.co/bEr7U7H3 #inchttp://t.co#irdependent: | | 3/29/2012 19:29 |
| kazydw | preciouslib | 3/29/2012 19:06 | RT @preciousliberty: Arianna Huffington: Tired of #Cbama rhet:http://t.co#Obama #citize | | 3/29/2012 19:06 |
| preciouslib | preciouslib | 3/29/2012 19:05 | Arianna Huffington: Tired of #Obama rhetoric that never deliveihttp://t.co#Obama #citize | | 3/29/2012 19:05 |
| preciouslib | preciouslib | 3/29/2012 19:21 | IRS data VERIFIES top 10% pay 70% of income taxes: Click 'Summhttp://t.co#p2 #1u #dem | | 3/29/2012 19:21 |
| preciouslib | preciouslib | 3/29/2012 19:40 | Obama's spending agenda is systematically dismantling our finahttp://t.co#irdependent: | | 3/29/2012 19:40 |
| truevineflo | preciouslib | 3/29/2012 19:29 | RT @preciousliberty: Arianna Huffington: Tired of #Cbama rhet:http://t.co#Obama #citize | | 3/29/2012 19:29 |
| penguinpor | rbpundit | 3/29/2012 19:38 | RT @JoleneAL: Classic! RT @MelissaTweets: RT @RBPundit: Fauhttp://t.co#tcot #p2 | | 3/29/2012 19:38 |
| penguinpor | melissatwe | 3/29/2012 19:38 | RT @JoleneAL: Classic! RT @MelissaTweets: RT @RBPundit: Fauhttp://t.co#tcot #p2 | | 3/29/2012 19:38 |
| wes_wheel | thegrudger | 3/29/2012 19:37 | @theGrudgeRetort @markos @evale72 How does it end Medicare exactly: #tcot #p2 | | 3/29/2012 19:37 |
| wes_wheel | thegrudger | 3/29/2012 19:37 | @theGrudgeRetort @markos @evale72 How does it end Medicare exactly: #tcot #p2 | | 3/29/2012 19:37 |
| thegrudger | wes_wheel | 3/29/2012 19:37 | RT @Wes_Wheeler: @theGrudgeRetort @markos @evale72 How does it e #tcot #p2 | | 3/29/2012 19:37 |
| clgold36 | coondawg6 | 3/29/2012 19:30 | RT @Coondawg68: Sarah Palin: I think the president's comment http://t.co#tcot #p2 | | 3/29/2012 19:30 |
| foolishrepc | qc16 | 3/29/2012 19:40 | RT @QC16: Spike Lee and the Democrats are a gang of terrorists. Truth hurt #tcot #p2 | | 3/29/2012 19:40 |
| thegrudger | thegrudger | 3/29/2012 19:09 | The dirty secret Dems don't want you to know: they couldn't care less abou #p2 #cot | | 3/29/2012 19:09 |
| thegrudger | thegrudger | 3/29/2012 19:41 | If you think the GOP is blocking jobs bills, you're a morcn. #p2 #cot | #p2 #cot | 3/29/2012 19:41 |
| thegrudger | thegrudger | 3/29/2012 19:41 | The Ryan budget is what a real "jobs bill" looks like. #p2 #cot | #p2 #cot | 3/29/2012 19:41 |
| ravenant01 | thegrudger | 3/29/2012 19:41 | RT @theGrudgeRetort: If you think the GOP is blocking jobs bills, you're a r #p2 #cot | | 3/29/2012 19:41 |
| politicaltod | politicaltod | 3/29/2012 19:14 | Sometimes I don't know if liberals are more irritating or amusing. I guess I' #tcot #p2 #p21 | | 3/29/2012 19:14 |
| truevineflo | truevineflo | 3/29/2012 19:45 | RT @Blueberrier0341: 2 years in a row Obama's budget doesn't get one sin #twisters #cot | | 3/29/2012 19:45 |
| sterlingvoti | sterlingvoti | 3/29/2012 19:06 | RT @PrcgressFlcrica: Is the conservative consensus crumbling chhttp://t.co#equality #ggt | | 3/29/2012 19:06 |
| cccupypclit | left4liberty | 3/29/2012 19:06 | RT @Left4Liberty: NDAA Lawsuit: Press Conference & Occupy Whhttp://t.co#L4L #p2 | | 3/29/2012 19:06 |
| cccupypclit | occupypoli | 3/29/2012 19:15 | Gwen Moore and the Violence Against Women Act #p2 http://t http://t.co#p2 | | 3/29/2012 19:15 |
| cccupypclit | torianbrow | 3/29/2012 19:18 | RT @TorianBrown: Top Congressman: Generals Are Lying to Me; http://t.co#tlot #p2 | | 3/29/2012 19:18 |
| cccupypclit | torianbrow | 3/29/2012 19:44 | RT @TorianBrown: US Fed judge: it might be unconstitutional tchhttp://t.co#tlot #p2 | | 3/29/2012 19:44 |

Edges

Vertices

Groups

Group Vertices

Overall Metrics

Twitter

| A | AC | AD | AE | AF | AG | AH | AI | AJ | AK |
|------------|----------|-----------|--------|-----------|------------------------------------------------------------------|----------------------------------|---------------------------------|-----------|--------------------------------|
| Vertex | Followed | Followers | Tweets | Favorites | Description | Location | Web | Time Zone | Time Zone UTC Offset (Seconds) |
| banksterb | 1 | 60 | 1600 | 3 | | | | | |
| serpentin | 6291 | 6523 | 38231 | 4 | Geologist. Gen trapped in GOP-land | Central T | | | -21600 |
| umtaha20 | 274 | 127 | 14496 | 331 | | | | | |
| occupywe | 10 | 144 | 45156 | 0 | #OccupyWallStreet: see these smart & highly effective alternativ | | | | |
| fcj316 | 286 | 151 | 7715 | 4 | Just another BrChicago, Illi | http://o | Central T | | -21600 |
| marin_ale | 301 | 406 | 2419 | 100 | I write for @PcNew York, N | http://g | Central T | | -21600 |
| freedom | 38 | 105 | 4026 | 1506 | | Bahrain | | Riyadh | 10800 |
| redline_1 | 546 | 339 | 31714 | 69 | | | | Quito | -18000 |
| echelonx | 1998 | 1670 | 8005 | 263 | Writer, Singer, Euless, TX | http://j | Central T | | -21600 |
| dcooktob | 5 | 1 | 1 | 0 | | | | | |
| this_just | 25 | 13 | 30 | 0 | if you don't knobakersfield, | http://w | Mountai | | -25200 |
| the_numt | 98 | 65 | 195 | 1 | More than just a number... | | | | |
| yousef172 | 204 | 24 | 2845 | 1992 | | | | | |
| maonati | 101 | 58 | 9291 | 14 | | | | | |
| robbyukul | 1935 | 385 | 740 | 41 | I'm #SingerSongwriter in # | http://soundcloud.com/robby-adam | | | |
| philolmst | 593 | 254 | 575 | 1 | Online Market | San Francisc | http://2 | Pacific T | -28800 |
| motsx120 | 52 | 7 | 137 | 14 | | | | Pacific T | -28800 |
| mzdm2 | 51 | 70 | 16813 | 8 | | | | | |
| averyoslo | 281 | 548 | 6585 | 51 | I'm a global citi | Netherland | http://averyoslo.wordpress.com/ | | |
| bnerino67 | 100 | 11 | 25 | 9 | Happily married to my wife of 24 years | | | | |
| revjim52 | 10 | 14 | 564 | 7 | | | | | |
| docpballe | 40 | 20 | 381 | 1 | | Bahrain | | | |
| slantedtin | 30 | 34 | 12 | 0 | Leaning right s | Seattle, WA | http://tl | Arizona | -25200 |
| totitru1 | 199 | 77 | 231 | 113 | | USA | | Eastern T | -18000 |
| paulpaulr | 14 | 3 | 20 | 0 | 76 year old. ret | SEATTLE. WASHINGTON. U.S.A. | | | |

Twitter Network Analysis

- **Vertex level metrics**
- Applying social network analysis to Twitter activity, users are characterized based on their connectivity in the network.
- Measuring how central users are in the network reveals the influential users and their connections to one another.
- ***In and out degree centrality***
- Degree centrality metrics count the number of connections (edges) a user (vertex) has in the network.
- Because Twitter networks are directed (e.g., @Aviva may mention @Hans, but @Hans may not mention @Aviva), degree centrality can take two forms.

Twitter Network Analysis

- ***In and out degree centrality***
- *In-degree centrality* measures the number of edges *others have initiated with a* vertex. For instance, if @Aviva was mentioned 5 times by users in a Twitter topic-network, her in-degree centrality metric would be 5.
- *Out-degree centrality counts the number of edge a vertex has initiated with others.* If @Hans mentioned 10 other users in his tweet, his out-degree centrality would be 10.

Twitter Network Analysis

- ***Betweenness centrality***
- Degree centrality metrics define a vertex's centrality by number of connections in a network.
- A user (i.e., a vertex) may also be central in a network because it connects users that would otherwise be disconnected or less connected. Betweenness centrality measures the extent to which a vertex plays this bridging role in a network.
- Specifically, betweenness centrality measures the extent that the user falls on the shortest path between other pairs of users in the network.
- The more people depend on a user to make connections with other people, the higher that user's betweenness centrality becomes.

Twitter Network Analysis

- ***User reciprocity***
- A relationship between two users is reciprocal, or mutual, if each user has initiated a tie with the other user
- Reciprocal relationships between individuals may indicate a wide range of social attributes, such as cooperation, trust, exchange of opinions, and power balance.
- At the user level, the Reciprocated Vertex Pair Ratio is measured as the number of users one is connected with (alters) that are reciprocal over the total number of alters.
- The portion of reciprocated relationships of the total number of relationships a vertex has with others in the network.

Twitter Network Analysis

- On Twitter, for example, a reciprocal or mutual relationship between two users can be established if they follow one another.
- If Aviva is connected with 10 other users on Twitter (whether following or being followed), and 5 of these users relationships are mutual (i.e., Aviva follows 5 users who also follow her), Aviva's reciprocity value will be 0.5.
- Reciprocity can be used to evaluate users' relationship building. When establishing a social media presence, users often aim to attract the attention of influential users by giving them attention (retweeting, posting hyperlinks, tagging, mentioning, etc.).
- Reciprocity metrics can be used to evaluate the success of this strategy. Reciprocity can also be measured for the entire network, or clusters in it.

Twitter Network Analysis

- In your dataset, sort the Reciprocated Vertex Pair Ratio column from largest to smallest.
 - What is the highest reciprocity ratio?
- A closer look will reveal that users with such a perfect reciprocity value, are not highly connected in the network.
- Find their in and out degree metrics. Pretty low, right? The reason is simple: the more connected users are in the network, the less likely they are to have all their connections reciprocated.
- There is typically an *overall negative correlation between between* users' degree (in or out) and their reciprocity values. The more connected users are, the lower their reciprocity value is *likely to be*. *It is therefore helpful to first find the top users in your network, and then compare and contrast their reciprocity ratios.*

Twitter Network Analysis

- **Network-level metrics**
- ***Overall metrics***
- Taking a social networks approach to data analysis shifts the focus from individual characteristics of users, to their connectivity-related characteristics.
- Another unique characteristic of social network analysis is the focus on metrics that describe a group of users in a connected component. On the Edges worksheet, an edge is the unit of analysis.
- On the Vertices worksheet, the vertex (the user) is the unit of analysis.
- In the Overall Metrics worksheet, the entire network is the unit of analysis.

Twitter Network Analysis

- **Network-level metrics**
- **Overall metrics**
- *Vertices.* The number of users in the Twitter search network.
- *Unique Edges.* The number of ties in the networks, excluding duplicates. For example, if @Joelle mentioned @Muhammad in two tweets within this network it will be counted as a single unique edge between @Joelle and @Muhammad.
- *Edges With Duplicates.* The number of duplicate relationships between users. For example, if @Joelle mentioned @Muhammad in two tweets within this network, this will be considered an edge with duplicates.
- *Total Edges.* The sum of Unique Edges and Edges With Duplicates.

Twitter Network Analysis

- **Network-level metrics**
- ***Overall metrics***
- *Self Loops.* In its original use, an edge is counted as a self loop when a user initiates a tie with itself.
- *Connected Components.* A component is a unit of one or more users that have connections among them. The Connected Components is a simple count of these components.
- *Single-Vertex Connected Components.* These are isolated users who are talking about the topic of your network, but in a given dataset, are not connected to others by an edge. In Twitter networks, these will be individuals who post a tweet that is not a mention, reply to, or retweet.

Twitter Network Analysis

- ***Graph density***
- Twitter networks vary in terms of their interconnectedness.
- Some networks are more tightly interconnected, by mentioning and replying to one another. In other networks, users are only sparsely connected, rarely mentioning or replying to others.
- Network density is measured as the number of possible or potential connections (i.e., edges), over the number of actual connections.
- Density values range between zero and one, and can be thought of as the percent of all possible edges that are realized.
- The calculation is a slightly different for directed and undirected networks, as directed networks have twice as many possible edge

Twitter Network Analysis

- ***Graph reciprocity***
- Reciprocity metrics at the user level were discussed earlier. At the network level, reciprocity measures the extent to which ties among a group of vertices are mutual.
- Reciprocity is measured as a proportion of mutual edges to the overall number of edges in a network.
- Values range between 0 (i.e., no mutual ties in the network) to 1 (i.e., all edge are mutual in the network). Other approaches to reciprocity metrics exist. Similarly to network density, reciprocity is associated with network size (number of vertices).
- The larger the network, the smaller the reciprocity are likely to be. Comparing networks in terms of their reciprocity, you should take into consideration the network size.

Twitter Network Analysis

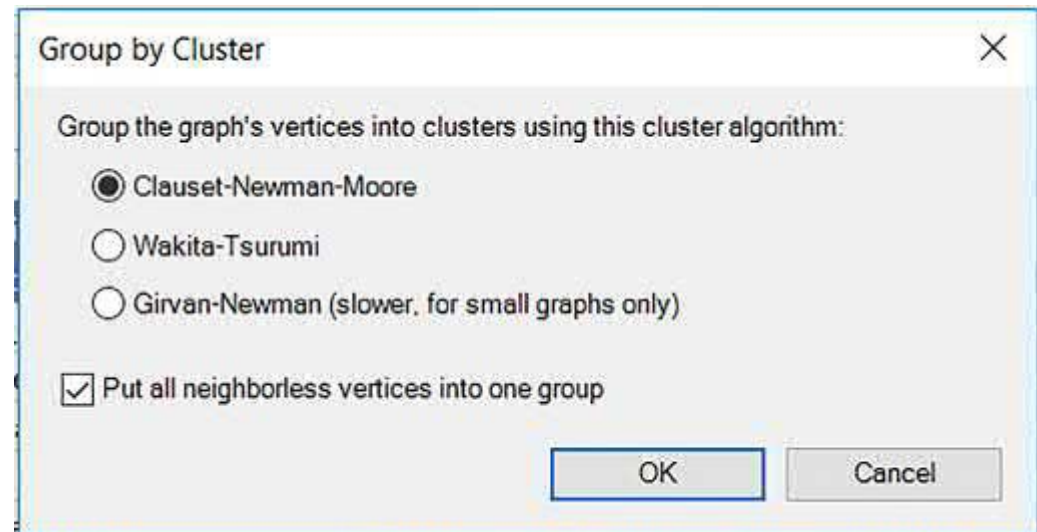
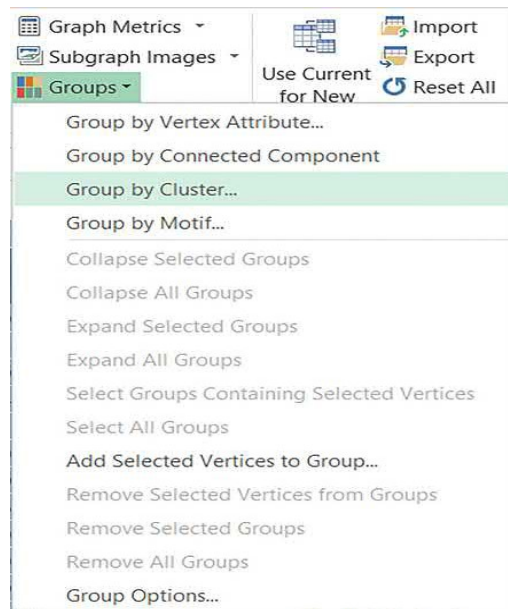
- **Groups**
- In social networks, smaller sub-groups of densely interconnected users – clusters – often arise. Clusters, also referred to as communities, refer to subgroups in a network in which vertices are substantially more connected to one another than to vertices outside that subgroup.
- In Twitter political networks, users' exposure to tweets derives from the users they follow.
- Twitter users, then, are more likely to read content posted by their cluster-mates than by users in other clusters.
- Likewise, users in a given cluster choose to expose themselves to the same set of hubs, which serve as popular information sources among these users.

Twitter Network Analysis

- **Groups**
- Clusters are identified by applying a mathematical algorithm that assigns vertices (i.e., users) to subgroups of relatively more connected groups of vertices in the network.
- The Clauset-Newman-Moore algorithm, used in NodeXL, enables you to analyze large network datasets to efficiently find subgroups.
- Identify clusters by following these steps:
 - In the Analysis section of the NodeXL Ribbon, click on Groups and select Group by Cluster...
 - In the pop up menu, select Clauset-Newman- Moore and check the Put all neighborless vertices into one group checkbox. The latter will create a “fake” cluster to accommodate all users who are not connected to others (isolates). Click OK.

Twitter Network Analysis

- **Groups**
- Identify clusters by following these steps:
 - A new worksheet will appear, called Groups.
 - To calculate group-level metrics: Click on Graph Metrics, then Deselect All, and then select Overall Graph Metrics and Group Metrics. Click Calculate Metrics.



Twitter Network Analysis

- **Visualization**
- User-level visual properties:
- We have already identified your top users. In particular, you have calculated users' in-degree and betweenness centrality values.
- Use the Autofill Columns tool to associate users' visual properties with centrality metrics

Twitter Network Analysis

Autofill Columns

Edges Vertices Groups

| When Autofill is clicked, fill in these worksheet columns... | ...using the values in these source columns, if the columns exist | Options |
|--------------------------------------------------------------|-------------------------------------------------------------------|----------------------------------|
| Vertex Color | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Shape | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Size | In-Degree | <input type="button" value="→"/> |
| Vertex Opacity | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Visibility | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Label | <input type="text"/> | <input type="button" value="→"/> |
| Fill Color | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Label Position | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Tooltip | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Layout Order | <input type="text"/> | <input type="button" value="→"/> |
| Vertex X | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Y | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Polar R | <input type="text"/> | <input type="button" value="→"/> |
| Vertex Polar Angle | <input type="text"/> | <input type="button" value="→"/> |

Clear All Worksheet Columns Now

Reset All Autofill Settings

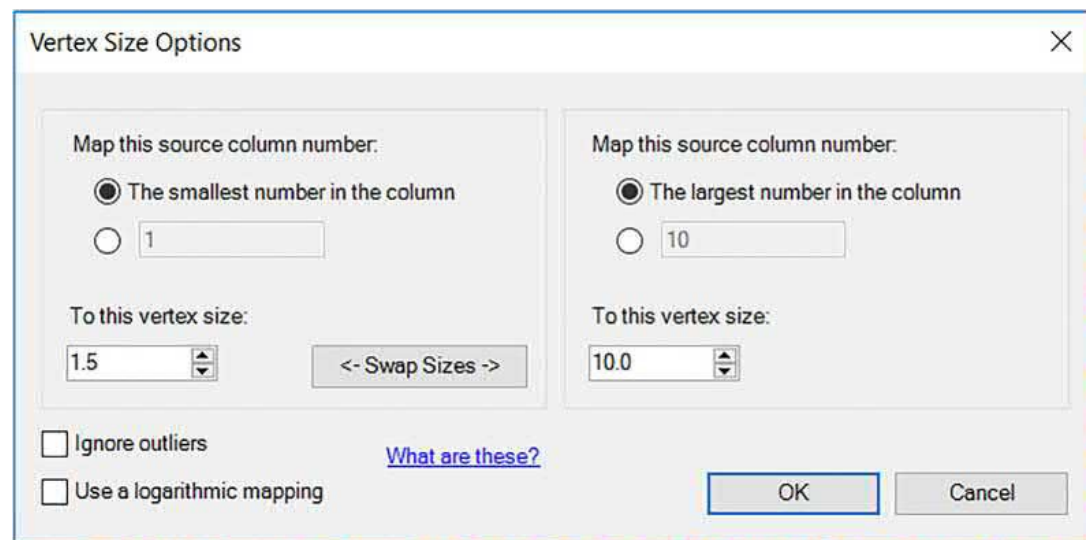
Autofill Close

Twitter Network Analysis

- Rather than simply identifying the key actors by their size, add labels to the key users. Because you are interested in changing users' visual properties, turn to the Vertices worksheet.
 - Find the In-Degree column and sort it by size, from large to small.
 - Find the Shape column and set it to the type Label for the users with the highest in-degree.
 - When selecting Label as the shape, NodeXL looks for the actual label in the corresponding cell in the Label column. This cell is currently empty. Type a label in it. A good label would be either the name of the user (e.g., NodeXL Project) or its Twitter handle (e.g., @nodexl).
 - Refresh the graph. Can you see the labels?
 - Change the size manually to something larger if needed. Refresh the graph.
 - Change the color. Just type in a major color in the corresponding cell in the Color column (e.g. Red). Refresh the graph.

Twitter Network Analysis

- Continue and assign labels and colors to all your top in-degree users. Be creative. You can select labels for all top in-degree users and select different colors, based on the identity of these users (e.g., news media, bloggers, politicians, etc.).
- You can also make the changes, by right-clicking on a vertex in the graph and selecting Edit selected vertex properties.
- You may notice that other nodes may cover the labeled nodes. To resolve this issue use the Autofill Columns dialog to set the Vertex Layout Order to In-Degree.



Twitter Network Analysis

- **Cluster-level layout and visual properties:**
- NodeXL provides an array of network layout options.
- It is often helpful to layout the network by groups, so clusters are highlighted.
- Earlier, you got NodeXL to disregard groups in the graph pane. Now, though, reversing this will let you better organize your graph layout:
 - Open the Group Options from the Groups dropdown menu on the NodeXL Ribbon.
 - Deselect Skip Groups.
 - You can decide where the visual properties (Shape and Color) for vertices come from, the Vertices or the Groups worksheets.

Twitter Network Analysis

- **Cluster-level layout and visual properties:**

- Keep the Colors coming from the Groups worksheet (the first option).
- Select the Vertices worksheet as the source of Shapes (the second option). This will allow you to keep the hub labeling. Next, layout the graph by clusters:
- In the Graph Pane click on the layout algorithm drop down menu.
- In the Layout style, select the second option: Lay out each of the graph's groups in its own box.

Layout Options

Margin: 6

Layout style

☐ Lay out the entire graph in the entire graph pane (typical case)

☒ Lay out each of the graph's groups in its own box

Box layout algorithm: Treemap

Width of the box outlines: 1

Intergroup edges: Show

☐ Use the Grid layout for groups that don't have many edges

☐ Lay out the graph's smaller connected components in boxes at the bottom of the graph pane

Maximum size of the connected components to lay out in boxes: 3 vertices

Box size: 16

Fruchterman-Reingold layout

Strength of the repulsive force between vertices: 3.0

Iterations per layout: 10

Reset All OK Cancel