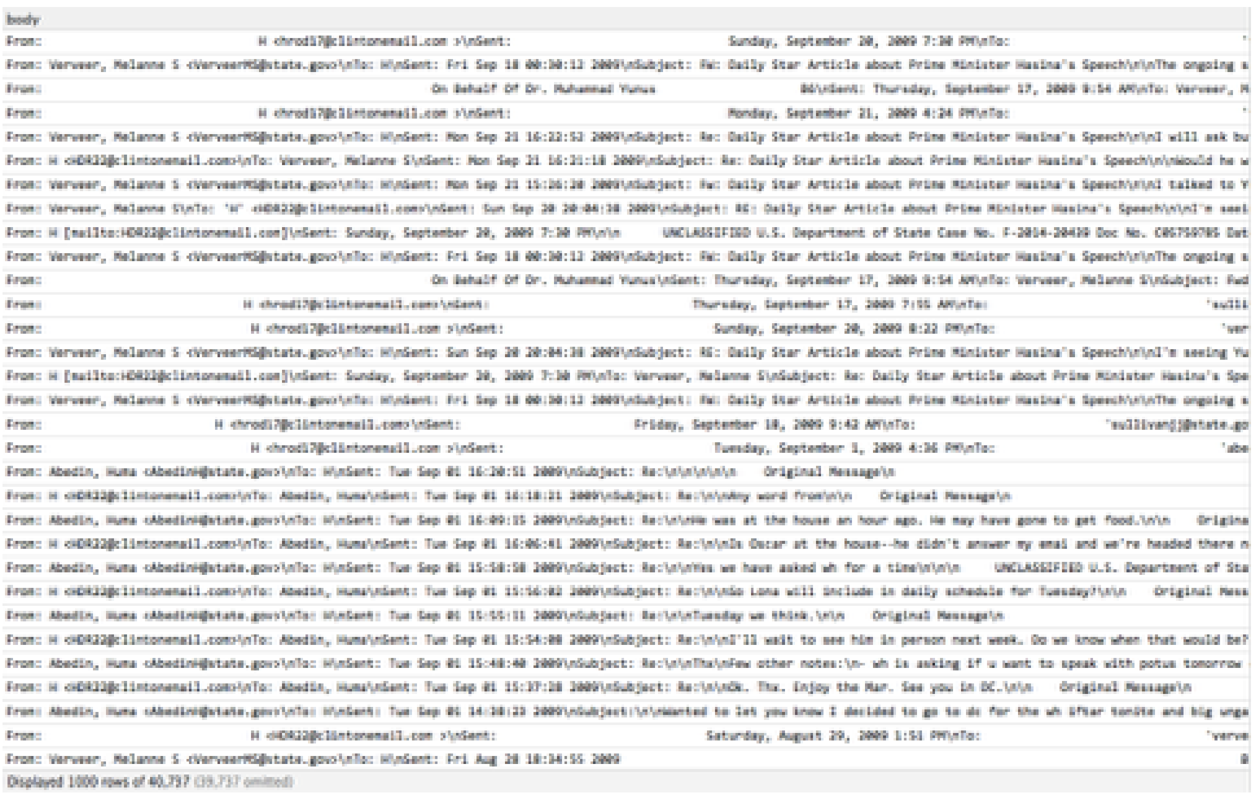
Lab report 2

Amir Imani - ai2335

27 April 2016

Describe the social network(s) to me, in terms of how it was collected, what it represents and so forth. Also give me basic topography of the network: the nature of the ties; direction of ties; overall density; and if attributes are with the network, the distribution of the categories and variables of those attributes.

I have downloaded the publicly available dataset of Hillary's emails. It is a SQL file containing more than 41K emails. Each row, is an email containing the body of email in addition to meta data such as sender, recipents, classification (image 1).



body
From: H <hrod17@clintonemail.com >\nSent: Sunday, September 20, 2009 7:30 PM\nTo:
From: Verveer, Melanne S <VerveerMS@state.gov>\nTo: H\nSent: Fri Sep 18 00:30:12 2009\nSubject: Fw: Daily Star Article about Prime Minister Hasina's Speech\n\nThe ongoing s
From: On Behalf Of Dr. Muhammad Yunus 0\nSent: Thursday, September 17, 2009 9:54 AM\nTo: Verveer, M
From: H <hrod17@clintonemail.com >\nSent: Monday, September 21, 2009 4:24 PM\nTo:
From: Verveer, Melanne S <VerveerMS@state.gov>\nTo: H\nSent: Mon Sep 21 16:22:52 2009\nSubject: Re: Daily Star Article about Prime Minister Hasina's Speech\n\nI will ask bu
From: H <HDR22@clintonemail.com>\nTo: Verveer, Melanne S\nSent: Mon Sep 21 16:21:18 2009\nSubject: Re: Daily Star Article about Prime Minister Hasina's Speech\n\nWould he w
From: Verveer, Melanne S <VerveerMS@state.gov>\nTo: H\nSent: Mon Sep 21 15:26:20 2009\nSubject: Fw: Daily Star Article about Prime Minister Hasina's Speech\n\nI talked to Y
From: Verveer, Melanne S\nTo: 'H' <HDR22@clintonemail.com>\nSent: Sun Sep 20 20:04:38 2009\nSubject: RE: Daily Star Article about Prime Minister Hasina's Speech\n\nI'm seei
From: H [mailto:HDR22@clintonemail.com]\nSent: Sunday, September 20, 2009 7:30 PM\n\n UNCLASSIFIED U.S. Department of State Case No. F-2014-20439 Doc No. C05759785 Dat
From: Verveer, Melanne S <VerveerMS@state.gov>\nTo: H\nSent: Fri Sep 18 00:30:12 2009\nSubject: Fw: Daily Star Article about Prime Minister Hasina's Speech\n\nThe ongoing s
From: On Behalf Of Dr. Muhammad Yunus\nSent: Thursday, September 17, 2009 9:54 AM\nTo: Verveer, Melanne S\nSubject: Fwd
From: H <hrod17@clintonemail.com>\nSent: Thursday, September 17, 2009 7:55 AM\nTo: 'sulli
From: H <hrod17@clintonemail.com >\nSent: Sunday, September 20, 2009 8:22 PM\nTo: 'ver
From: Verveer, Melanne S <VerveerMS@state.gov>\nTo: H\nSent: Sun Sep 20 20:04:38 2009\nSubject: RE: Daily Star Article about Prime Minister Hasina's Speech\n\nI'm seeing Yu
From: H [mailto:HDR22@clintonemail.com]\nSent: Sunday, September 20, 2009 7:30 PM\nTo: Verveer, Melanne S\nSubject: Re: Daily Star Article about Prime Minister Hasina's Spe
From: Verveer, Melanne S <VerveerMS@state.gov>\nTo: H\nSent: Fri Sep 18 00:30:12 2009\nSubject: Fw: Daily Star Article about Prime Minister Hasina's Speech\n\nThe ongoing s
From: H <hrod17@clintonemail.com>\nSent: Friday, September 18, 2009 9:42 AM\nTo: 'sullivanjj@state.go
From: H <hrod17@clintonemail.com >\nSent: Tuesday, September 1, 2009 4:36 PM\nTo: 'abe
From: Abedin, Huma <AbedinH@state.gov>\nTo: H\nSent: Tue Sep 01 16:20:51 2009\nSubject: Re:\n\n\n\n\n Original Message\n
From: H <HDR22@clintonemail.com>\nTo: Abedin, Huma\nSent: Tue Sep 01 16:18:21 2009\nSubject: Re:\n\nAny word from\n\n Original Message\n
From: Abedin, Huma <AbedinH@state.gov>\nTo: H\nSent: Tue Sep 01 16:09:35 2009\nSubject: Re:\n\nHe was at the house an hour ago. He may have gone to get food.\n\n Origina
From: H <HDR22@clintonemail.com>\nTo: Abedin, Huma\nSent: Tue Sep 01 16:06:41 2009\nSubject: Re:\n\nIs Oscar at the house--he didn't answer my email and we're headed there n
From: Abedin, Huma <AbedinH@state.gov>\nTo: H\nSent: Tue Sep 01 15:58:58 2009\nSubject: Re:\n\nYes we have asked wh for a time\n\n\n UNCLASSIFIED U.S. Department of Sta
From: H <HDR22@clintonemail.com>\nTo: Abedin, Huma\nSent: Tue Sep 01 15:56:02 2009\nSubject: Re:\n\nSo Lona will include in daily schedule for Tuesday?\n\n Original Mess
From: Abedin, Huma <AbedinH@state.gov>\nTo: H\nSent: Tue Sep 01 15:55:11 2009\nSubject: Re:\n\nTuesday we think.\n\n Original Message\n
From: H <HDR22@clintonemail.com>\nTo: Abedin, Huma\nSent: Tue Sep 01 15:54:08 2009\nSubject: Re:\n\nI'll wait to see him in person next week. Do we know when that would be?
From: Abedin, Huma <AbedinH@state.gov>\nTo: H\nSent: Tue Sep 01 15:48:40 2009\nSubject: Re:\n\nThx\nFew other notes:\n- wh is asking if u want to speak with potus tomorrow
From: H <HDR22@clintonemail.com>\nTo: Abedin, Huma\nSent: Tue Sep 01 15:37:28 2009\nSubject: Re:\n\nOk. Thx. Enjoy the Mar. See you in DC.\n\n Original Message\n
From: Abedin, Huma <AbedinH@state.gov>\nTo: H\nSent: Tue Sep 01 14:38:23 2009\nSubject:\n\nWanted to let you know I decided to go to dc for the wh iftar tonite and big unga
From: H <HDR22@clintonemail.com >\nSent: Saturday, August 29, 2009 1:51 PM\nTo: 'verve
From: Verveer, Melanne S <VerveerMS@state.gov>\nTo: H\nSent: Fri Aug 28 18:34:55 2009

Displayed 1000 rows of 40,757 (39,757 omitted)

image 1- Initial data

For the purpose of this lab, I wanted to construct the Hillary's correspondence network. To do so, I extracted senders and receivers to use as a directed edge list using patterns in the data set and regular expressions. After having a dataframe with two columns of "sender" and "receivers", I cleaned the data more by removing special characters, white spaces, etc.

An interesting challenge of this process was entity resolution as not only there were a lot of typos in email addresses (for example 'Jake Sullivan' was typed 'Lake Sullivan') and there were variation of one name (like 'Human Abedin' and 'Abedin Huma'), but also single persons had different names (for example H, HRD, HClinton, etc for Hillary). To address that, I carried entity resolution using Levenshtein distance for similarity between words, and also decision trees for a higher resolution.

After cleaning up the data frame, there were more than 20,000 names. Therefore, I decided to construct the network based on the top 100 names with the highest number of emails (aggregated by both sent and received). Also in the case the more than one person was in the recipients field, I picked only the first one assuming that the first one in the list is more important in the communication.

Finally, I subset the data based on if the email was classified or not to create two seperate networks calling them 'classified' and 'unclassified' graphs. Ties in both networks are directed and denote if node A has sent an email to node B.

Basic topography of both networks are shown below.

Network	Number of Nodes	Number of Edges	Density
Classified	1236	1947	0.001274468
Unclassified	195	292	0.007679158

table 1. basic topography of networks

---

2. Calculate degree centrality; closeness centrality; betweenness centrality; and eigenvector centrality. Correlate those measures of centrality. Highlight which nodes are most central and least central, along different dimensions.

a summary of all centrality measures is shown in the table below.

```
##                inDegreeC outDegreeC totalDegreeC inClosenessC outClosenessC
## h                     64        604          668 7.062780e-07  1.185508e-05
## cherylmills           64        266          330 7.062556e-07  1.179579e-05
## humaabedin            50        150          200 7.063025e-07  1.166875e-05
## sullivanjacob         56        104          160 7.063005e-07  1.177315e-05
## opsnewsticker          0         44           44 6.551105e-07  6.787871e-07
## mchalejuditha          8         30           38 7.062870e-07  1.154748e-05
##                totalClosenessC betweennessC       vector
## h                 1.185508e-05    55138.428 1.000000e+00
## cherylmills       1.179579e-05    26819.782 5.011265e-01
## humaabedin        1.166875e-05    30551.407 6.258795e-01
## sullivanjacob     1.177315e-05    29365.378 5.334829e-01
## opsnewsticker     6.787871e-07        0.000 3.683675e-08
## mchalejuditha     1.154748e-05     5000.559 5.370628e-02
```

As expected, **Hillary** has the highest degree centrality. She also has the highest betweeness and eignevector centrality. These are all intuitive as it is her email network and every single email has her either as the sender or recipients. Huma Abedin(Hillary's aide), Cheryil Mills (Cheif of staff), and Jacob Sullivan (Policy advisor) who have the highest degree after hillary also have other highest centrality measures.

correlation between measures of centrality for the classified network.

```
##                  inDegreeC    outDegreeC totalDegreeC  inClosenessC
## inDegreeC       1.00000000  0.8168445365  0.865819706  0.0289936334
## outDegreeC      0.81684454  1.0000000000  0.995874457 -0.0006531362
## totalDegreeC    0.86581971  0.9958744568  1.000000000  0.0039942812
```

```
## inClosenessC    0.02899363 -0.0006531362  0.003994281  1.0000000000
## outClosenessC   0.42574328  0.2861668409  0.315186859 -0.0426309717
## totalClosenessC 0.42574328  0.2861668409  0.315186859 -0.0426309717
## betweennessC    0.85405218  0.8441790737  0.866571088  0.0197498369
## vector          0.90203863  0.9132646369  0.934043095  0.0124298811
##                 outClosenessC totalClosenessC betweennessC     vector
## inDegreeC          0.42574328      0.42574328   0.85405218 0.90203863
## outDegreeC         0.28616684      0.28616684   0.84417907 0.91326464
## totalDegreeC       0.31518686      0.31518686   0.86657109 0.93404310
## inClosenessC      -0.04263097     -0.04263097   0.01974984 0.01242988
## outClosenessC      1.00000000      1.00000000   0.37945055 0.29677608
## totalClosenessC    1.00000000      1.00000000   0.37945055 0.29677608
## betweennessC       0.37945055      0.37945055   1.00000000 0.87151240
## vector             0.29677608      0.29677608   0.87151240 1.00000000
```

> **3b. If you don't have a network with attribute data, then pick another network to compare your first network against. Calculate all of the same measures as above for Network #2. Consider if normalization is appropriate for any of these measures. Then state some hypothesis about why some (or all of the) measures of centrality in one network will be the same or different from the second network. Explain why you think these two networks should be similar or different.**

As explained earlier, the other network I picked is the correspondance network for unclassified emails. a summary of all centrality measures is shown in the table below.

```
##                 inDegreeU outDegreeU totalDegreeU inClosenessU
## h                      16         64           80 3.496137e-05
## cherylmills            23         22           45 3.505943e-05
## sullivanjacob          23         21           44 3.505328e-05
## burnswilliam            8         13           21 3.499195e-05
## humaabedin              6         14           20 3.475239e-05
## feltmanjeffreyd         4         15           19 3.501768e-05
##                 outClosenessU totalClosenessU betweennessU    vector
## h                0.0001541782    0.0001541782    3990.8333 1.0000000
## cherylmills      0.0001514463    0.0001514463    3591.1667 0.3315094
## sullivanjacob    0.0001538935    0.0001538935    5231.3333 0.7750269
## burnswilliam     0.0001407658    0.0001407658     890.1667 0.1138866
## humaabedin       0.0001495663    0.0001495663     658.0000 0.4983735
## feltmanjeffreyd  0.0001517451    0.0001517451    4103.5000 0.1261271
```

Again Hillary has the highest degree followed by the same people in the classified network.

correlation between measures of centrality for the classified network.

```
##                 inDegreeU outDegreeU totalDegreeU inClosenessU
## inDegreeU       1.0000000 0.75041889    0.8794767   0.17529604
## outDegreeU      0.7504189 1.00000000    0.9745557   0.01613123
## totalDegreeU    0.8794767 0.97455572    1.0000000   0.07106200
## inClosenessU    0.1752960 0.01613123    0.0710620   1.00000000
## outClosenessU   0.2843117 0.45899250    0.4269242  -0.35748351
## totalClosenessU 0.2843117 0.45899250    0.4269242  -0.35748351
```

```
## betweennessU    0.8477590 0.78593752    0.8534248   0.08448205
## vector          0.8250938 0.88866129    0.9197073   0.07221896
##                 outClosenessU totalClosenessU betweennessU     vector
## inDegreeU           0.2843117       0.2843117   0.84775897 0.82509378
## outDegreeU          0.4589925       0.4589925   0.78593752 0.88866129
## totalDegreeU        0.4269242       0.4269242   0.85342484 0.91970734
## inClosenessU       -0.3574835      -0.3574835   0.08448205 0.07221896
## outClosenessU       1.0000000       1.0000000   0.37802132 0.32182815
## totalClosenessU     1.0000000       1.0000000   0.37802132 0.32182815
## betweennessU        0.3780213       0.3780213   1.00000000 0.77001411
## vector              0.3218282       0.3218282   0.77001411 1.00000000
```

---

4. In either case, when you are done above, then considers alternate specifications of your variables and codings and decisions and models. What would you want to consider changing and why. If you can, report on what are the consequences of those changes?

One observation from the previous tables is that becuase the number of emails to/from hillary is much more than the others, it skews the result for the less frequent nodes. Therefore, an alternative can be removing her from the network and calculating same centrality network measures for the remaining network.

doing so, the basic network topography comparing to the previous ones will be:

Network	Number of Nodes	Number of Edges	Density
Classified	1236	1947	0.001274468
Unclassified	195	292	0.007679158
Hillary-less	732	1280	0.002388844

centrality measures will also look like

```
##                 inDegreeH outDegreeH totalDegreeH inClosenessH
## cherylmills            63        265          328 2.106194e-06
## humaabedin             49        149          198 2.106580e-06
## sullivanjacob          55        103          158 2.106545e-06
## opsnewsticker           0         44           44 1.868838e-06
## mchalejuditha           7         29           36 2.106483e-06
## reinesphilippei        12         23           35 2.106336e-06
##                 outClosenessH totalClosenessH betweennessH       vector
## cherylmills      1.546360e-05    1.546360e-05    20748.806 5.776256e-06
## humaabedin       1.536594e-05    1.536594e-05    19706.066 1.024929e-06
## sullivanjacob    1.542353e-05    1.542353e-05    19634.194 1.217215e-06
## opsnewsticker    1.984033e-06    1.984033e-06        0.000 1.000000e+00
## mchalejuditha    1.528071e-05    1.528071e-05     4894.021 8.687026e-08
## reinesphilippei  1.529450e-05    1.529450e-05     4320.300 2.700383e-07
```

Another interesting measure in this context is number and size of cliques. with Hillary in the network, it will be a large clique with 1236 nodes in it. However, removing her from the network, there is 732 clique with 6 nodes as the largest and ofcourse all of them include Cheryl Mills, Huma Abedin, and Jake Sullivan. It worth exploring more who are the members of each clique and what is their association.