VirginiaTech
Institute for Critical Technology and Applied Science

Ted and Karyn
Hume Center for National Security and Technology

Data Mining Academic Emails to Model Employee Behaviors and Analyze Organizational Structure

Kayla Straub

Master's Defense

3/25/2016

hume@vt.edu
www.hume.vt.edu

Outline

VirginiaTech

- Problem statement and contributions
- Email dataset
- Feature extraction
- Algorithm design
- Performance analysis
- Future work and conclusions

3/25/2016 2

Problem Overview

VirginiaTech

- Email is everywhere!
- Difficult to research email because of inherent privacy concerns
- Lack of modern email datasets with accurate job title labels
- What information about an organization is embedded in the organizational email communication?
- Organic vs. official organizational charts

What information about an organization can be extracted from emails?

3/25/2016 3

Applications

VirginiaTech

- Overall methods can be applied to any communication system
  - Cell phone, website links, social media, network connections
- This particular type of analysis could benefit:
  - In Dec. 2015, GE completed downsize and merger of subsidiary General Electric Capital Corporation
  - On Jan. 1 2016, Northrop Grumman combined two of its four business sectors, Electronic Systems and Information Systems
  - In late 2016, the merger of two major chemical companies: DuPont and Dow will be finalized before splitting into three new companies

3/25/2016 4

Contributions

VirginiaTech

- Present a new email dataset based on academic emails of the center
- Job title classification results that outperform previous work
- A method to automatically generate organic hierarchy from analyzing emails
- Paper submitted to IJCAI 2016

Improved email analysis feature set and classification results

3/25/2016 5

VirginiaTech
Institute for Critical Technology and Applied Science

Ted and Karyn
Hume Center for National Security and Technology

Email Dataset

3/25/2016 6

Enron Dataset

- Benchmark dataset for email analysis
- Released in 2004
- Used for research into spam classification, email categorization, and recipient prediction
- Issues with the dataset
  - 99.99% overlap between emails sent as "CC" and those sent as "BCC"
  - Some emails addresses, folders, and names are misspelled
  - Inconsistent email address formats make mapping to employees difficult
- Issues with job title labels
  - No labels for 29 employees
  - Clear mislabeling errors for at least 4 employees

3/25/2016 7

Prior Work

- Prior work in hierarchy analysis often uses the text of the emails with natural language processing features, mainly on Enron
- Historically, email analysis without text uses two types of features:
  - Traffic-based: statistical features based on single emails
  - Social-based: features calculated based on email interactions between people
- Success in determining community structures has been found using the two types separately
  - Namata et al. 2006 used traffic-based features to predict Enron job titles
  - Wilson and Banzhaf 2009 found Enron's important groups from strictly social features
- Rowe et al. 2005 used a combination of features to automatically construct the Enron social hierarchy

3/25/2016 8

Hume Email Data Collection

- Worked with Virginia Tech's Internal Review Board (IRB) to approve data collection procedures and privacy concerns
  - All subject and body text was hashed using MD5 algorithm
  - Data collection process was performed using automated scripts
  - No identifying information is revealed in analysis
  - All data stored on secure, password-protected Hume Center server
- Hashing example:

Employment Opportunity at Doosan Fuel Cell America → b4dabd5884ecd175283065dc605c2172

3/25/2016 9

Email Parsing Process

VirginiaTech

- Challenge: Email formats are inconsistent
  - Forwards are expressed as “Fw:” or “Fwd:” or “FW:”
  - Email address encoded with Unicode
  - Some emails have HTML- needed to identify, then parse

- Process:
  - Write a python script to extract data
  - Test on personal emails to find inconsistent formatting issues
  - Ran script on mail server and saved all email metadata into MySQL database

3/25/2016 10

Collected Data From Each Email

VirginiaTech

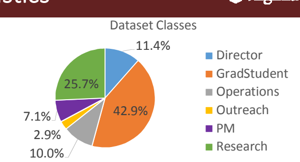
- Destination and source email address
- Email time stamp
- Subject prefix (e.g., Re:, Fwd:)
- Hash of subject after removing prefix
- Hash of body text
- Length of subject in characters
- Length of body text in characters
- Number of attachments
- Indicator if email was digitally signed
- Indicator if email was encrypted

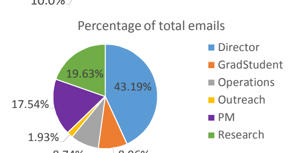
3/25/2016 11

Dataset Description and Statistics

VirginiaTech

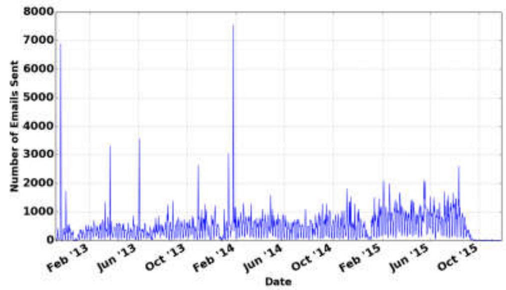
- 37 volunteers in the study
  - Plus 32 additional employees from the data
- 585,096 emails over 3 years
- 6 job categories:
  - Director
  - Graduate Student
  - Operations
  - Outreach
  - Project Management (PM)
  - Research





3/25/2016 12

Hume Email Visualization



Hume Center vs. Enron Dataset Comparison

- Hume Center dataset:
  - More modern
  - More distinct emails
  - Longer time period
  - Academic emails
- Enron dataset:
  - More employees
  - More distinct email addresses
  - Corporate emails

	Hume Center	Enron
Time	11/2012-11/2015	1/2000-9/2002
Distinct Email Addresses	32,118	75,406
Participants	37	158
Distinct Emails	585,096	252,759

VirginiaTech Institute for Critical Technology and Applied Science | Ted and Karyn Hume Center for National Security and Technology

Feature Extraction

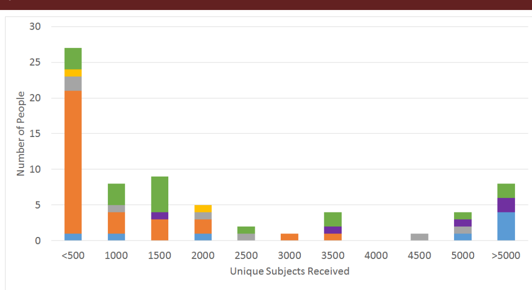
Features

- Features quantify information extracted from the email metadata
- Two categories:
  - Traffic-based – 84 features
  - Social-based – 30 features
- Total features: 114
- Calculated using bash, MySQL, and python scripts
- Used as input to the machine learning algorithm

Traffic-Based Features

- Generated directly from the collected metadata
- Example raw features:
  - Unique subjects received
  - Number of signed emails received
  - Number of emails received as carbon copies
  - Average number of emails received per day
  - Number of emails sent after normal business hours
  - Number of emails sent within VT
  - Number of emails sent within Hume
- Also converted raw features as percentages

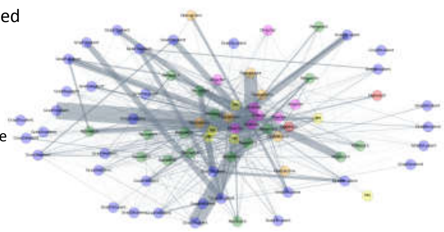
Example Traffic-based Feature



Social Graph

VirginiaTech

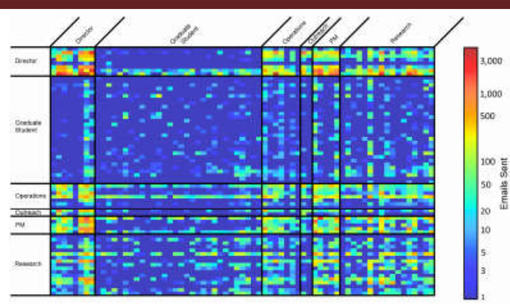
- Nodes represent people
- Edges are directed and represent emails exchanged between people
- Two different graphs used
  - Full graph draws edge between people who exchanged at least one email
  - 2532 edges, 86 nodes
  - Partial graph draws edge between people who exchanged at least 10 emails
  - 1319 edges, 82 nodes



3/25/2016 19

Alternative Social View - Adjacency Matrix

VirginiaTech



3/25/2016 20

Social-Based Features

VirginiaTech

- Degree measures
  - Degree of a node
  - Average neighbor degree
- Cliques
- Clustering Metrics
- Search Engine Algorithms
- Centrality Measures
- All features calculated for both the full and partial graph

3/25/2016 21

Betweenness Centrality

VirginiaTech

- There exists a shortest path between any node s and any other node t
- Betweenness centrality of a node i is the percentage of all shortest paths in graph $\mathcal{G}$ that traverse node i:

$$C_B(i) = \sum_{s,t \in \mathcal{V}} \frac{\sigma(s,t|i)}{\sigma(s,t)}$$

$\mathcal{V}$ is the set of all nodes in $\mathcal{G}$

$\sigma(s,t)$ is the number of shortest paths between s and t

$\sigma(s,t|i)$ is the number of those paths that pass through i

3/25/2016 22

Closeness Centrality

VirginiaTech

- Normalized inverse of the sum of shortest path distances from node i to all other nodes in the graph:

$$C(i) = \left(\frac{\sum_{j=1}^{n-1} d(i,j)}{n \quad 1} \right)^{-1}$$

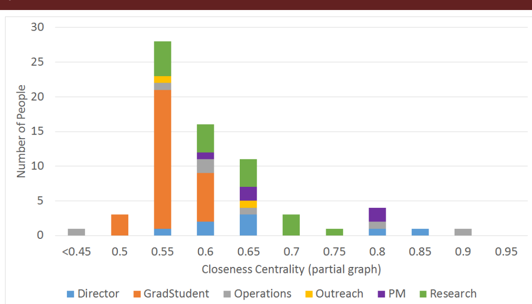
n is the number of nodes in graph $\mathcal{G}$

$d(i,j)$ is the minimum shortest path distance between node i and node j

3/25/2016 23

Example Social Feature

VirginiaTech



3/25/2016 24

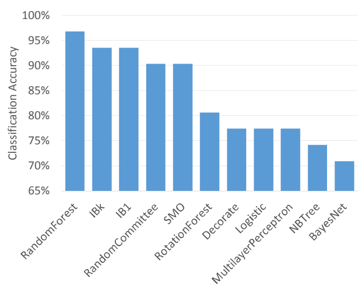
Ted and Karyn
Hume Center for National Security and Technology

VirginiaTech
Institute for Critical Technology and Applied Science

Algorithm Design

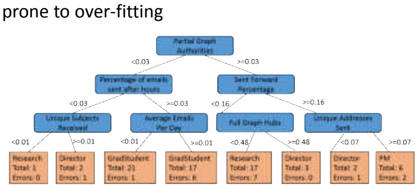
Algorithm Selection

- Used limited number of employees:
  - 29 participants
  - Removed people with less than 100 emails
  - Removed the two outreach employees
- Randomly split emails into train and test sets
- Using Weka, evaluated classification accuracy of several algorithms with default parameters



Random Trees

- Used to learn a series of rules for classification
- Learned using a greedy heuristic
  - Starting at the top, split on the best feature
  - Automatic feature selection
- Tree-based classifiers are prone to over-fitting
  - Low bias, high variance



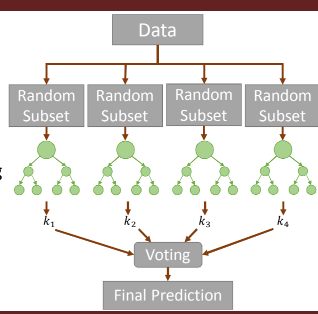
Determining the Best Split

- Entropy: Amount of randomness in the class distribution
  - $H(\text{Class})$
- Conditional Entropy: Amount of randomness in the class distribution when the attribute value is known
  - $H(\text{Class}|\text{Attribute})$
- Mutual Information:
  - $I(\text{Class}; \text{Attribute}) = H(\text{Class}) \quad H(\text{Class}|\text{Attribute})$
- Split on maximum mutual information:
  - $X \;\; = \arg\max_x I(\text{Class}; X) = \arg\max_x \; H(\text{Class}) \quad H(\text{Class}|X) = \arg\min_x H(\text{Class}|X)$
- Because variables are continuous, use thresholds to form discrete levels
  - $t \;\; = \arg\min_t H(\text{Class}|t) = \arg\min_t H(\text{Class}|X < t)P(X < t) + H(\text{Class}|X \geq t)P(X \geq t)$

3/25/2016 28

Random Forests

- Random forests are an ensemble method that is robust to overfitting
- Learn many deep random trees and take majority vote of prediction outputs
  - "Bagging", or Bootstrap Aggregating
- Random elements introduced in each tree:
  - Only a subset of features used
  - Only a subset of training data used



3/25/2016 29

Parameters

- Data split:
  - 35% training, 30% cross-validation, 35% testing
- Number of trees
  - 750
- Number of features considered per branch split
  - 7
  - 6% of total possible features
- Number of samples used per tree
  - $\frac{2N}{3}$ observations, with replacement

3/25/2016 30

Feature Selection

New	Feature	Type	Mutual Information
	Unique subjects received	Traffic	0.728
✓	**Total signed emails received**	**Traffic**	**0.728**
	Hubs (partial graph)	Social	0.589
	Number of emails received as forwards	Traffic	0.519
	Current flow closeness centrality (partial graph)	Social	0.512
	Pagerank (partial graph)	Social	0.512
	Percentage of emails sent with unique addresses	Traffic	0.51
	Number of emails received as carbon copies	Traffic	0.5
	Pagerank (full graph)	Social	0.492
	Average number of emails received per day	Traffic	0.489
✓	**Communicability betweenness centrality (partial graph)**	**Social**	**0.486**
	Communicability centrality (partial graph)	Social	0.486

3/25/2016 31

VirginiaTech
Institute for Critical Technology and Applied Science

Ted and Karyn
Hume Center for National Security and Technology

Performance Analysis

3/25/2016 32

Classification Results

- Training: random 35% of emails
- Cross-validation: random 30% of emails
- Testing: random 35% of emails
- Using Random Forests and tuned parameters

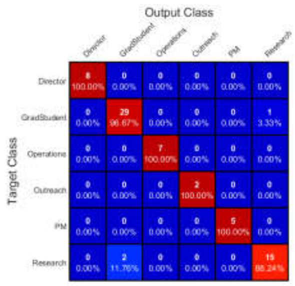
- Potential bias:
  - Training and testing performed on the same people due to small sample size

3/25/2016 33

Classification Results

VirginiaTech

- Overall very accurate: 95.7%
- Confusion between graduate students and researchers
- These errors are understandable
  - Some doctoral students have been working at the center for 3-5 years
  - Some research faculty are also graduate students



3/25/2016 34

Behavior Over Time

VirginiaTech

- Considered employees with at least 10 months of emails
- Procedure:
  - Data split into 1-month segments
  - First month used as training
  - Months 2-10 used as test data
- Confirms assumption that email behavior is constant in time

3/25/2016 35

Leave-One-Out Cross-Validation

VirginiaTech

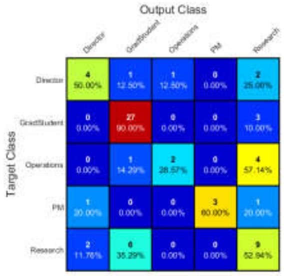
- Concerned about bias from training and testing on same people
- Procedure
  - Train on all but one sample
  - Test on that sample
  - Repeat for all samples
- Removed Outreach for this test because only contains two samples

3/25/2016 36

LOOCV Classification Results

VirginiaTech

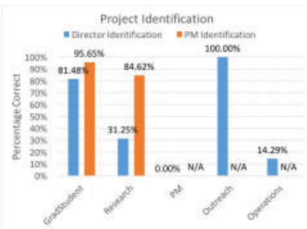
- Lower accuracy than previous result
  - 67.2% overall
  - Due to low sample size
  - Different roles in each group
- Graduate students very accurate
  - 90.0%
- In each category, there are more true positives than false positives from any other label
- Namata et al. 2006 performed same analysis on Enron with 62.09% accuracy
  - Improvement of 5.11%



Hierarchy Analysis

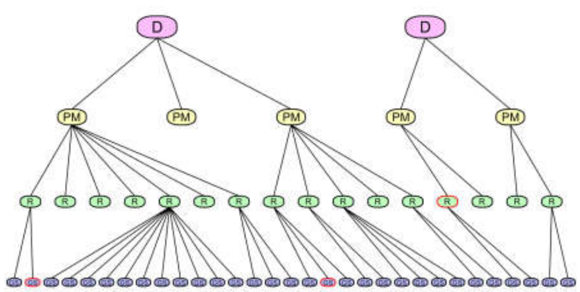
VirginiaTech

- Gathered ground-truth project labels
  - Director for all classes
  - PM for graduate students and researchers
- Project labels were unavailable for some past employees
- Overall:
  - 52.63% communicated most with assigned director
  - 91.67% communicated most with assigned PM
- Overall PM relationships are more consistent with official hierarchy than director relationships



Algorithm-Generated Organic Org Chart

VirginiaTech

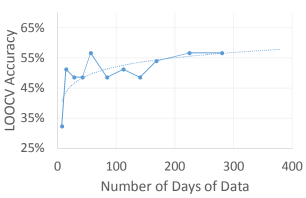


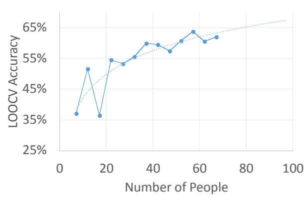
Ted and Karyn Hume Center for National Security and Technology

VirginiaTech Institute for Critical Technology and Applied Science

Conclusions and Future Work

More Data would Improve





Future Work and Deep Learning Applications

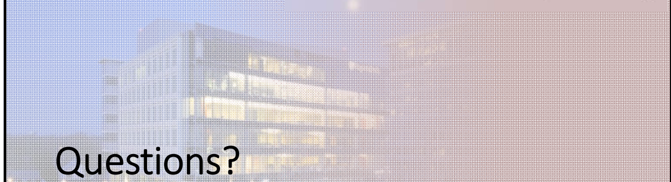
- Apply to larger datasets and/or different types of data
  - Cleaned Enron corpus
  - Hillary Clinton Email Dataset
  - Twitter dataset
- Investigate process of releasing the fully anonymized dataset
- Potential Deep Learning Application
  - More sophisticated algorithms could be used
    - But need much more semi-supervised data
  - Deep Belief Networks
  - Unsupervised training
  - Supervised back propagation

Conclusions

- Created a brand new email dataset from raw emails
  - With accurate job title labels
  - Approximately the size of Enron, but with fewer people
  - Privacy precautions
- This dataset is meant to be representative of data any company could collect without violating the privacy of their employees
- Highly accurate classification results based on historical data
  - Showed that email behavior is constant with time
- Small dataset lead to low LOOCV accuracy
  - Improved on previous Enron result
- An organic organization chart was produced that represented the email relationships of the center

3/25/2016 43

VirginiaTech Institute for Critical Technology and Applied Science | Ted and Karyn Hume Center for National Security and Technology



Questions?

3/25/2016 44

Thank You

- Committee: Dr. McGwier, Dr. Beex, Dr. Buehrer, Dr. Huang
- Dr. Ernst and Dr. Headley
- Faculty, staff, and graduate students of the Hume Center

3/25/2016 45

VirginiaTech
Institute for Critical Technology and Applied Science

Ted and Karyn
Hume Center for National Security and Technology

Backup

HTML Example

```
</head>
<body dir=3D"ltr" style=3D"font-size:12pt;color:#000000;background-color:#F=
FFFFF;font-family:Calibri,Arial,Helvetica,sans-serif;">
<p><br>
</p>
<p>FYI, attached is information on a reliability theory course offered by D=
r. Joel Nachlas in the ISE Department.<br>
</p>
<div>
<p><br>
</p>
<div id=3D"Signature">
<div class=3D"BodyFragment"><font size=3D"2">
<div class=3D"PlainText">Leslie K. Pendleton, Ph.D.<br>
Director, Student Services<br>
Department of Electrical and Computer Engineering (Mail Code 0111)<br>
Virginia Tech<br>
340 Whittemore Hall<br>
Blacksburg, VA 24061<br>
USA<br>
Email: pendleton@vt.edu<br>
Phone: (540) 231-8219<br>
Fax: (540) 231-3362<br>
<br>
</div>
</font></div>
</div>
</div>
<p></p>
<p><br>
</p>
<p><br>
</p>
</body>
</html>
```


All Features

total_sent
unique_addresses_sent
unique_add_sent_perc
unique_subjects_sent
unique_sub_sent_perc
total_received
unique_addresses_received
uniqueadd_rec_perc
unique_subjects_received
unique_sub_rec_perc
total_emails
total_sent_signed
total_sent_signed_perc
unique_addresses_sent_signed
unique_add_sent_perc_signed
unique_subjects_sent_signed
unique_sub_sent_perc_signed
total_received_signed
total_rec_signed_perc
unique_addresses_received_signed
unique_add_rec_perc_signed
unique_subjects_received_signed
unique_sub_rec_perc_signed
total_sent_encrypted
total_sent_encrypted_perc
unique_addresses_sent_encrypted
unique_add_sent_perc_encrypted
unique_subjects_sent_encrypted
unique_sub_sent_perc_encrypted
total_received_encrypted
total_rec_encrypted_perc
unique_addresses_received_encrypted
unique_add_rec_perc_encrypted
unique_subjects_received_encrypted
unique_sub_rec_perc_encrypted
inter_hume_sent
inter_hume_received
inter_vt_sent
inter_vt_received
sent_to
sent_to_perc
sent_cc
sent_cc_perc
rec_to
rec_to_perc
rec_cc
rec_cc_perc
avg_recipients_sent
avg_recipients_rec
avg_body_chars_sent
avg_body_chars_rec
var_body_chars_sent
var_body_chars_rec
after_hours_sent
after_hours_sent_perc
after_hours_rec
after_hours_rec_perc

All Features

VirginiaTech

after_hours_sent_hume
after_hours_sent_hume_perc
after_hours_rec_hume
after_hours_rec_hume_perc
avg_sent_per_day
avg_rec_per_day
avg_emails_per_day
attached_sent
attached_rec
avg_attachments_sent
avg_attachments_rec
sent_re
sent_re_perc
sent_fw
sent_fw_perc
rec_re
rec_re_perc
rec_fw
rec_fw_perc
avg_subject_chars_sent
avg_subject_chars_rec
var_subject_chars_sent
var_subject_chars_rec
fg_between_centrality
pg_between_centrality
pg_avg_neighbor_degree
fg_avg_neighbor_degree
pg_clustering
fg_clustering
pg_closeness_centrality
fg_closeness_centrality
pg_degree_centrality
fg_degree_centrality
pg_current_flow_closeness_centrality
fg_current_flow_closeness_centrality
pg_current_flow_betweenness_centrality
fg_current_flow_betweenness_centrality
pg_communicability_centrality
fg_communicability_centrality
pg_communicability_betweenness_centrality
fg_communicability_betweenness_centrality
pg_load_centrality
fg_load_centrality
pg_square_clustering
fg_square_clustering
fg_eccentricity
pg_eccentricity
pg_pagerank
fg_pagerank
pg_hubs
pg_authorities
fg_hubs
fg_authorities
pg_avg_shortest_paths
fg_avg_shortest_paths
pg_num_cliques
fg_num_cliques

Neighborhood Degree

VirginiaTech

- The neighborhood of node i is comprised of all nodes that are connected to i via edges.
- The average neighbor degree is therefore

$$k_{avg,i} = \frac{1}{|N(i)|} \sum_{j \in N(i)} k_j$$

- $|N(i)|$ is the number of neighbors of node i
- k_j is the degree of node j

Triangle Clustering

VirginiaTech

- Compares the number of triangles node i is a part of to the maximum number of possible triangles.

$$C_{3,i} = \frac{2}{k_i(k_i \quad 1)} \sum_{m,n} \left(\widetilde{w}_{i,m} \widetilde{w}_{m,n} \widetilde{w}_{n,i}\right)^{\frac{1}{3}}$$

- If node i has degree k_i, there can be at most $\frac{k_i(k_i-1)}{2}$ triangles involving i
- Normalize edge weights compared to maximum: $\widetilde{w}_{i,m} = \frac{w_{i,m}}{\max(w_{i,m})}$, then take geometric mean

Degree Centrality

VirginiaTech

- Degree centrality of a node i is the percentage of nodes within the graph that are connected to node i

$$C_{d,i} = \frac{k_i}{n \quad 1}$$

Communicability Centrality

VirginiaTech

- Also known as subgraph centrality
- Consider all closed walks in graph $\mathcal{G}$ of length k
- Of those walks, those that begin on node i are denoted as $\mu_k(i)$
- The communicability centrality of node i is:

$$SC(i) = \sum_{k=1}^{\infty} \frac{\mu_k(i)}{k!}$$

Some Worst Performing Features

VirginiaTech

New	Feature	Type	Mutual Information
✓	**Total Received Encrypted**	Traffic	0
✓	**Unique Addresses Sent Signed**	Traffic	0
	Inter-Hume Received	Traffic	0
	Unique Addresses Received	Traffic	0
✓	**Total Sent Encrypted**	Traffic	0
	Number of Cliques (full graph)	Social	0
✓	**Inter-VT Sent**	Traffic	0
	Betweenness Centrality (partial graph)	Social	0
	Clustering (partial graph)	Social	0
	Average Neighbor Degree (full graph)	Social	0
	Average Subject Characters Received	Traffic	0