Analytics Practicum Midterm Report

TEAM 17: APA

Aayush Garg, Prekshaa Uppin, Akshita Dhandhania

Contents

Introduction

Human Resource Analytics is the idea of using data in the organizational context to understand different factors about employees such as their degree of collaboration and influence. Collaboration, being a crucial part of managing an organization is a valuable determinant in understanding how decisions are made and how relations are built. Furthermore, influence can provide a blueprint of the hubs of information flow and effective change in the organization. Through this project, we aim to provide a way of comprehending these factors through deep data analysis and patterns observed in communication interactions (email and instant messaging) of employees.

TrustSphere is a market leader in Relationship Analytics, delivering solutions through Sales Analytics, Risk Analytics and People Analytics. Their goal is to help clients find the value of their associated networks for improving key business challenges such as sales force effectiveness, enterprise-wide collaboration, participation and contribution statistics and corporate governance.

Objective

1**. Perform Feature Engineering to create a new 'Trust Score' algorithm**: A trust score is an aggregate weightage that shows the strength of communication tie between two employees in a social network.

2. **Develop a dashboard that displays various metrics** that would quantify the collaboration between employees, identify the most influential employees and give managers a high-level view of these statistics to maintain a collaborative and efficient workplace.

3. **Research and validate** the potential of a **Hybrid Centrality** (potentially a combination of betweeness and degree) calculated from email communication data as a measure of influence score.

Data

1. Email Data

Columns	Column Explanation
Date	Date of the E-mail
Remote IP	If the email exchange is external, then this column shows the external person's email
Remote	The trust sphere employee who is receiving the email
Remote Domain	Always trustsphere
Local	E-mail address of the person sending the email
Local Domain	Domain of the person who is sending the email
Originator	Inbound, Outbound or Internal
Direction	Always trustsphere in this case
Domain Group	Email Header (Subject Line)
Subject	Type of message: email/im/voice/sms
Inbound Count	Number of emails received
Outbound count	Number of emails sent
Size	Size of the message
Msgid	Encoded Message ID

2. Staff Data

Columns	Column Explanation
Name	Name of the employee
Hierarchy	Designation of the employee
Department	Department of the employee
Location	The location where the employee is based

Data Statistics and Cleaning

1. Email Data

1. Data extracted from **11/26/2016 8:00 am** to **02/01/2017 00:00 am**
2. Before Cleaning:
   a) 14 columns of data
   b) 121,154 rows of data
3. Cleaning Steps:
   a) Remove emails with Subject not equal to 'email'
      i) Rationale: Analysis only on email data
   b) Remove emails with Originator not equal to 'internal'
      i) Rationale: Analysis only on internal communication
   c) Removing System Emails from Local and Remote:
      i) Rationale: Analysis on collaboration between real employees only
      ii) List of system emails found in Local:
         (1) accounting@trustsphere.com (1)
         (2) amazons3@trustsphere.com (3)
         (3) analytics@trustsphere.com (134)
         (4) careers@trustsphere.com (4)
         (5) customer.care@trustsphere.com (120)
         (6) heartbeat@trustsphere.com (1658)
         (7) info@trustsphere.com (1)
         (8) jira@trustsphere.com (1386)
         (9) marketing.team@trustsphere.com (322)
         (10)marketing@trustsphere.com (56)
         (11)northamericanteamcallactionitems@trustsphere.com (9)
         (12)peopleanalytics@trustsphere.com (175)
         (13)postman@trustsphere.com (1097)
         (14)postmaster@trustsphere.com (39)
         (15)sfdc@trustsphere.com (899)
         (16)sg.boardroom@trustsphere.com (22)
         (17)support@trustsphere.com (604)
         (18)trustsphere.office@trustsphere.com (15)
         (19)trustvault.selfservice@trustsphere.com (95)
         (20)tv.reports@trustsphere.com (1394)
         (21)wordpress@trustsphere.com (21)
         (22)zabbix@trustsphere.com (1739)
      iii) List of system emails found in Remote:
         (1) alerts.support@trustsphere.com (1)
         (2) aolia@intradyn.com (1)
         (3) crm.report@trustsphere.com (6197)
         (4) customer.care@trustsphere.com (5)

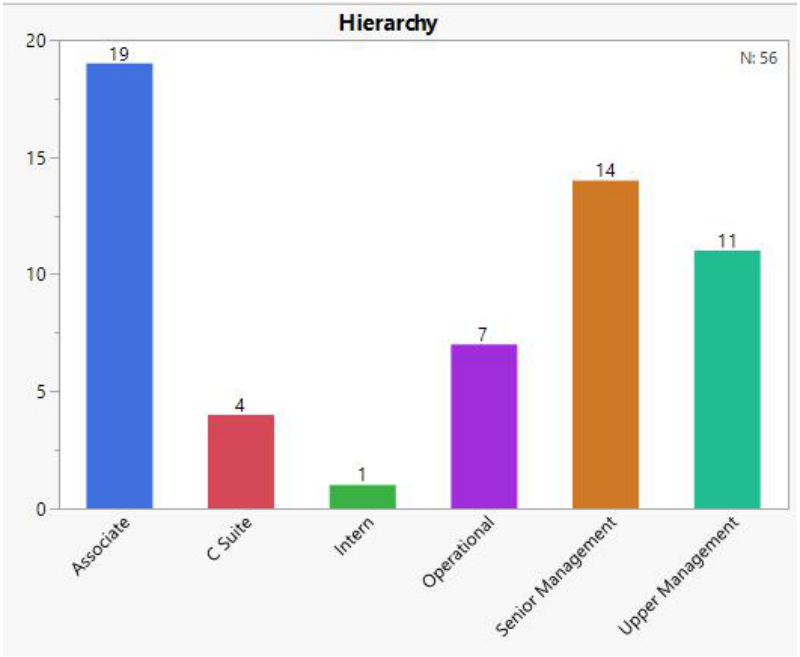
(5) dhartzler@intradyn.com (1)
(6) marketingteam@trustsphere.com (1)
(7) mgillard@intradyn.com (1)
(8) postman@trustsphere.com (20)
(9) sg.boardroom@trustsphere.com (25)
(10)sys.admin@trustsphere.com (12)

d) Remove unnecessary columns such as:
   i) Remote IP
   ii) Remote Domain
   iii) Local Domain
   iv) Direction
   v) Inbound count
   vi) Outbound count
   vii) Subject

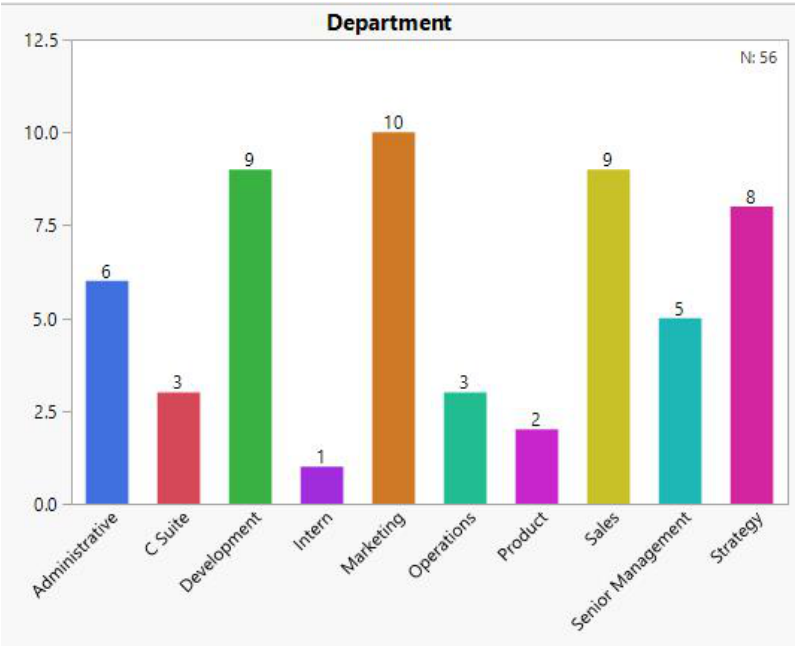
4. After Cleaning:
   a) 29,797 rows of data
   b) No missing data instance

2. Staff Data

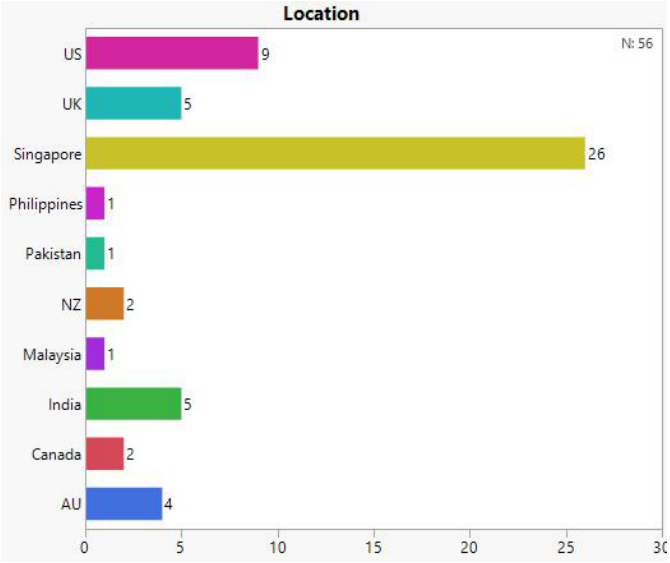
- 56 rows of data (56 Employees)
- 6 different Hierarchy Levels



- 10 different departments



- 10 different locations – Majority in Singapore



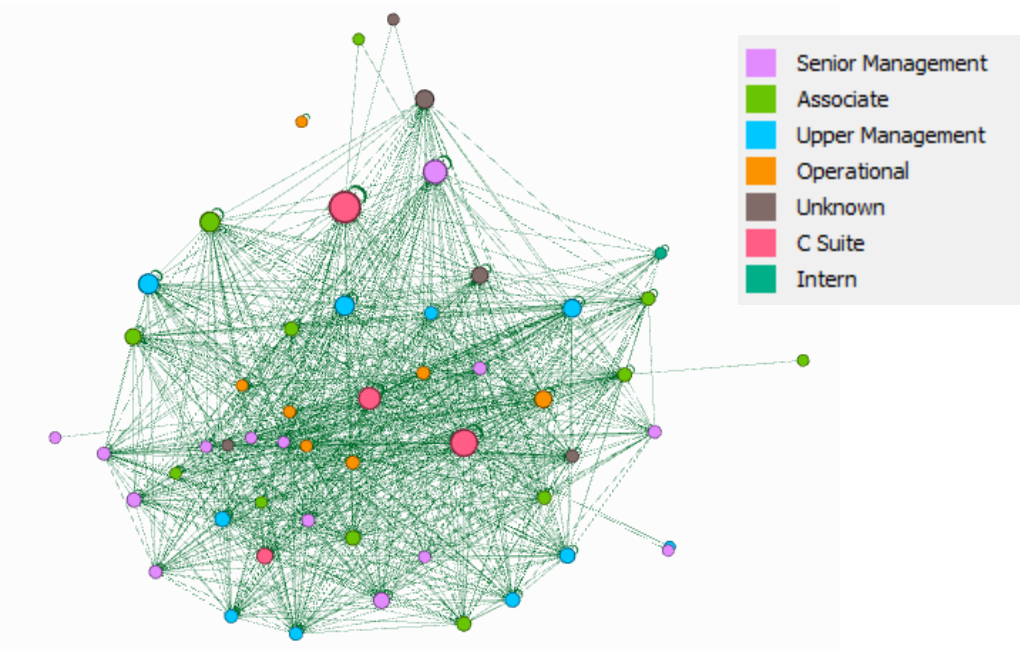
Data Exploration

Exploration: Email Data vs Staff Data

- **Highlighted in Pink:**
  - Employee is present in staff list but do not have any email interaction in the past 10 weeks.
  - There are 8 such employees.
  - These employees have probably left the company and staff data will be updated accordingly.
- **Highlighted in Yellow:**
  - Employee is present in the email interaction data for the past 10 weeks, but not present in the staff list.
  - There are 5 such employees.
  - These employees are probably new hires and staff data will be updates accordingly.

From Email Data (Last 10 weeks)		Staff List			
id		**Name**	**Hierarchy**	**Department**	**Location**
adesh.goel@trustsphere.com		Adesh Goel	C Suite	C Suite	Singapore
alistair.weatherill@trustsphere.com		Ajay Rana	Associate	Sales	India
amanda@trustsphere.com		Alistair Weatherill	Upper Management	Operations	UK
		Ananya Deshpande	Associate	Strategy	Singapore
annabel.koh@trustsphere.com		Annabel Koh	Associate	Strategy	Singapore
antony.ebelle@trustsphere.com					
aravind.mp@trustsphere.com					
		Anoshia Naseer	Associate	Strategy	Pakistan
arun.sundar@trustsphere.com		Arun Sundar	C Suite	Senior Management	Singapore
bersileus.sacamay@trustsphere.com					
brian.lebahn@trustsphere.com		Brian Lebahn	Upper Management	Senior Management	US
bryan.acedo@trustsphere.com		Bryan Acedo	Senior Management	Development	Philippines
dawn.radecki@trustsphere.com		Dawn Radecki	Upper Management	Marketing	US
deokant.pagasi@trustsphere.com					
dev.menon@trustsphere.com		Dev Menon	Upper Management	Sales	Singapore
		Elizabeth Botes	Upper Management	Marketing	US
esther.tan@trustsphere.com		Esther Tan	Operational	Administrative	Singapore
gabrielle.tremblay@trustsphere.com		Gabrielle Tremblay	Intern	Intern	Singapore

Exploration: Network



Node: Each employee

Node Color: Hierarchy

Node Size: Eigenvector Centrality

- No weights for edges – purely based on quantity
- Many Senior Management and Upper Management Employees seem to have a low centrality score.
- Possibly a biased solution
- **Inference**: Need for feature engineering to add weight that removes the bias

Methodology

The following is our methodology:

1. Understand the Scope of the project
2. Explore and clean the data
3. This step has 2 parts which happen in parallel:
    a. Perform Feature Engineering on email data
    b. Create and send out a Survey to create test data that would validate the hybrid centrality metric (influential score)
4. Using the features, create an aggregate score (new Trust Score) that will be used as weight for the communication network
5. Design suitable metrics that can be calculated from the weighted network
6. Test multiple hybrid centrality equations against survey results and finalize the algorithm
7. Develop a dashboard to display all metrics using R
8. Deliver to Client.

Feature Engineering

As-is Trust Score

Trustsphere's as-is trust score is currently based on high level features such as:

- Volume of emails sent and received
- When was the last interaction? (recency)
- Reply rate

Since these features are not accounting for various factors like variety and quality of interaction, Our team decided to create new features that would create a more representative trust score.

New features

The features we are using are:

1. Subject Line Weightage (for **quality** of information exchanged)
2. Email exchange ratio (for **frequency** of information exchanged)
3. Average Email Exchange Size (for **quantity** of information exchanged)
4. Email Chain Ratio (for **variety** of information exchanged)
5. Rate of exchange (for **regularity** of information exchanged)

Subject Line Weightage

We used SAS Enterprise Miner for text mining. We first performed text parsing and then text filtering on the subject line of the emails. Once we got the most frequently occurring terms along with their respective IDF (log) weightage, we decided to take an inverse of IDF, that is, 1/IDF as the actual weightage of the terms. We decided to do so because we wanted to give terms with a higher frequency, a higher weightage as we observed that these terms indeed had greater importance in the context of business email exchange. Thus, we finally took an inverse of IDF to weight the terms (as shown in the yellow bordered box below). We then chose the top 100 business related terms and gave a weightage to each row in our email exchange data, based on the occurrence of these terms in the subject line as shown below in the table. This helps us find important business related email exchanges.

Term	1/weight
meeting	0.200965
sugar	0.200361
update	0.187758
poc	0.177022
discussion	0.169808
opportunity	0.168067
sales	0.163212

Date	Target	Source	Originator	Domain group	Size	Msgid	Subject Weightage
1/31/2017 23:58	arun.sundar@trustsphere.com	hana.owens@trustsphere.com	internal	Call with HCLI	21358	<AM4PR0201MB1	0
1/31/2017 23:52	steve.allam@trustsphere.com	tom.butler@trustsphere.com	internal	RE: SOW_DBS_hk(2)-1 (003).docx	872031	<VI1PR0202MB29	0.138419167
1/31/2017 23:52	dev.menon@trustsphere.com	tom.butler@trustsphere.com	internal	RE: SOW_DBS_hk(2)-1 (003).docx	872031	<VI1PR0202MB29	0.138419167
1/31/2017 23:52	annabel.koh@trustsphere.com	tom.butler@trustsphere.com	internal	RE: SOW_DBS_hk(2)-1 (003).docx	872031	<VI1PR0202MB29	0.138419167
1/31/2017 23:52	arun.sundar@trustsphere.com	tom.butler@trustsphere.com	internal	RE: SOW_DBS_hk(2)-1 (003).docx	872031	<VI1PR0202MB29	0.138419167
1/31/2017 23:44	shaun.keating@trustsphere.com	warren.tait@trustsphere.com	internal	Re: Resignation	50644	<336252EC-0E96-	0
1/31/2017 23:42	warren.tait@trustsphere.com	shaun.keating@trustsphere.com	internal	Re: Resignation	42933	<C2C21FCF-6298-	0
1/31/2017 23:41	shaun.keating@trustsphere.com	warren.tait@trustsphere.com	internal	Resignation	19411	<DA248A80-1B76-	0
	rk.padginton@trustsphere.com	warren.tait@trustsphere.com	internal	Resignation	19411	<DA248A80-1B76-	0
	an.lebahn@trustsphere.com	manish.goel@trustsphere.com	internal	Re: meet with Graco next week	220240	<ed14fb28-4444-	0.137540144
	ha.chopra@trustsphere.com	mark.padginton@trustsphere.com	internal	FW: Sales Ready Lead > +40	70289	<A47EE504-69F4-	0.142603932
	un.keating@trustsphere.com	mark.padginton@trustsphere.com	internal	FW: Sales Ready Lead > +40	70289	<A47EE504-69F4-	0.142603932
	ya.bagga@trustsphere.com	mark.padginton@trustsphere.com	internal	FW: Sales Ready Lead > +40	70289	<A47EE504-69F4-	0.142603932
	nish.goel@trustsphere.com	brian.lebahn@trustsphere.com	internal	RE: meet with Graco next week	156697	<HE1PR0202MB26	0.137540144
	un.keating@trustsphere.com	warren.tait@trustsphere.com	internal	FW: Sales Ready Lead > +40	23870	<E2BA850F-40BA-	0.142603932
	rk.padginton@trustsphere.com	warren.tait@trustsphere.com	internal	FW: Sales Ready Lead > +40	23870	<E2BA850F-40BA-	0.142603932
	un.keating@trustsphere.com	manish.goel@trustsphere.com	internal	Re: IM screen	184126	<2427ee97-1109-	0
	vn.radecki@trustsphere.com	manish.goel@trustsphere.com	internal	Re: Sales Ready Lead > +40	120900	<985545a8-223d-	0.142603932
	n.sundar@trustsphere.com	manish.goel@trustsphere.com	internal	Re: Manish - approval pls	93906	<b68ba518-9fce-	0
1/31/2017 23:32	adesh.goel@trustsphere.com	manish.goel@trustsphere.com	internal	Re: Manish - approval pls	93906	<b68ba518-9fce-	0
1/31/2017 23:29	manish.goel@trustsphere.com	arun.sundar@trustsphere.com	internal	Re: Manish - approval pls	92621	<228053DF-F3C1-	0
1/31/2017 23:29	adesh.goel@trustsphere.com	arun.sundar@trustsphere.com	internal	Re: Manish - approval pls	92621	<228053DF-F3C1-	0
1/31/2017 23:29	arun.sundar@trustsphere.com	manish.goel@trustsphere.com	internal	Re: YC list	54081	<292dc4be-8416-	0.125580811
1/31/2017 23:28	arun.sundar@trustsphere.com	manish.goel@trustsphere.com	internal	Re: Manish - approval pls	110991	<70e83838-8ecf-e	0
1/31/2017 23:28	arun.sundar@trustsphere.com	manish.goel@trustsphere.com	internal	Re: Manish - approval pls	110343	<70e83838-8ecf-e	0
1/31/2017 23:28	adesh.goel@trustsphere.com	manish.goel@trustsphere.com	internal	Re: Manish - approval pls	110343	<70e83838-8ecf-e	0
1/31/2017 23:27	arun.sundar@trustsphere.com	manish.goel@trustsphere.com	internal	Re: Recommended Partner in UK?	46480	<d00af210-05e9-1	0.132450331
1/31/2017 23:20	shaun.keating@trustsphere.com	tom.butler@trustsphere.com	internal	RE: New release candidate for Sugar v1.2 - (manif	148279	<VI1PR0202MB29	0.166980592

Email Exchange Ratio

We wanted to check the number of emails exchanged between two employees to show how much they interact, collaborate and share information. We did this using the formula:

$$\frac{N_{ab}}{N_a + N_b - N_{ab}}$$ where

N_{ab}: Number of emails exchanged between A and B

N_a: Number of emails sent by A

N_b: Number of emails sent by B

The SQL statement used to apply this formula is shown in the figure below:

```
select t1.employee1, t1.employee2, (t1.total/(IFNULL(t2.total,0) + IFNULL(t3.total,0) - t1.total)) as
EmailExchangedRatio from

(select least(`Remote`,`Local`) as employee1, greatest(`Remote`,`Local`) as employee2, count(*) as total from
midterm group by least(`Remote`,`Local`), greatest(`Remote`,`Local`)) as t1

left join

(select `Local` as employee, count(*) as total from midterm group by `Local`) as t2

on t2.employee = t1.employee1

left join

(select `Local` as employee, count(*) as total from midterm group by `Local`) as t3

on t3.employee = t1.employee2
```

A screenshot of the results is shown below:

employee1	employee2	EmailExchangedRatio
adesh.goel@trustsphere.com	adesh.goel@trustsphere.com	0.0021
adesh.goel@trustsphere.com	alistair.weatherill@trustsphere.com	0.0710
adesh.goel@trustsphere.com	annabel.koh@trustsphere.com	0.0076
adesh.goel@trustsphere.com	antony.ebelle@trustsphere.com	0.0277
adesh.goel@trustsphere.com	aravind.mp@trustsphere.com	0.0042
adesh.goel@trustsphere.com	arun.sundar@trustsphere.com	0.1065
adesh.goel@trustsphere.com	bersileus.sacamay@trustsphere.com	0.0014
adesh.goel@trustsphere.com	brian.lebahn@trustsphere.com	0.0369
adesh.goel@trustsphere.com	bryan.acedo@trustsphere.com	0.0082
adesh.goel@trustsphere.com	dawn.radecki@trustsphere.com	0.0230

Average Email Exchange Size

Assuming that a larger email size shows larger amount of information exchange, we used the following formula to calculate the average email exchange size:

$$Average\ (Size\ of\ all\ emails\ exchanged\ between\ A\ and\ B)$$

The SQL statement used to apply this formula is shown in the figure below:

```
select least(`Remote`,`Local`) as employee1, greatest(`Remote`,`Local`) as employee2, avg(size)
from midterm group by least(`Remote`,`Local`), greatest(`Remote`,`Local`)
```

A screenshot of the results is shown below:

employee1	employee2	avg(size)
adesh.goel@trustsphere.com	adesh.goel@trustsphere.com	55786.3333
adesh.goel@trustsphere.com	alistair.weatherill@trustsphere.com	188445.8462
adesh.goel@trustsphere.com	annabel.koh@trustsphere.com	1191951.0714
adesh.goel@trustsphere.com	antony.ebelle@trustsphere.com	258309.8182
adesh.goel@trustsphere.com	aravind.mp@trustsphere.com	767172.2000
adesh.goel@trustsphere.com	arun.sundar@trustsphere.com	253348.0875
adesh.goel@trustsphere.com	bersileus.sacamay@trustsphere.com	44036.0000
adesh.goel@trustsphere.com	brian.lebahn@trustsphere.com	149990.2553
adesh.goel@trustsphere.com	bryan.acedo@trustsphere.com	23118.1111
adesh.goel@trustsphere.com	dawn.radecki@trustsphere.com	482241.5641
adesh.goel@trustsphere.com	dev.menon@trustsphere.com	38610.7544
adesh.goel@trustsphere.com	gabrielle.tremblay@trustsphere.com	44036.0000
adesh.goel@trustsphere.com	gladys.opone@trustsphere.com	32002.6667

Email Chain Ratio

Number of emails with unique subject lines shows number of different conversations taking place between employees. To capture this, we used the following formula:

$$\frac{N_u}{N_{ab}}$$

Where

N_u: Number of emails exchanged between A and B with unique subject lines

N_{ab}: Number of emails exchanged between A and B

The SQL statement used to apply this formula is shown in the figure below:

```
SELECT t1.employee1, t1.employee2, (t1.uniqueEmails/t1.total) as ratio from (SELECT least(`Remote`,`Local`)
as employee1, greatest(`Remote`,`Local`) as employee2, count(*) as total, count( distinct `Domain group`) as
uniqueEmails FROM midterm group by least(`Remote`,`Local`), greatest(`Remote`,`Local`)) as t1
```

A screenshot of the results is shown below:

employee1	employee2	ratio
adesh.goel@trustsphere.com	adesh.goel@trustsphere.com	1.0000
adesh.goel@trustsphere.com	alistair.weatherill@trustsphere.com	0.6795
adesh.goel@trustsphere.com	annabel.koh@trustsphere.com	0.6429
adesh.goel@trustsphere.com	antony.ebelle@trustsphere.com	0.6591
adesh.goel@trustsphere.com	aravind.mp@trustsphere.com	1.0000
adesh.goel@trustsphere.com	arun.sundar@trustsphere.com	0.8222
adesh.goel@trustsphere.com	bersileus.sacamay@trustsphere.com	1.0000
adesh.goel@trustsphere.com	brian.lebahn@trustsphere.com	0.8085
adesh.goel@trustsphere.com	bryan.acedo@trustsphere.com	1.0000
adesh.goel@trustsphere.com	dawn.radecki@trustsphere.com	0.6923

Rate of exchange of emails

Rate of exchange of emails shows how regularly employees interact with one another. Thus we used the following formula:

$$\frac{N_{ab}}{c}$$

Where

N_{ab}: Number of emails exchanged between A and B

c: number of weeks

The SQL statement used to apply this formula is shown in the figure below:

```
(select least(`Remote`,`Local`) as employee1, greatest(`Remote`,`Local`) as employee2, (count(*)/10) as total
from midterm group by least(`Remote`,`Local`), greatest(`Remote`,`Local`))
```

A screenshot of the results is shown below:

employee1	employee2	total
adesh.goel@trustsphere.com	adesh.goel@trustsphere.com	0.3000
adesh.goel@trustsphere.com	alistair.weatherill@trustsphere.com	7.8000
adesh.goel@trustsphere.com	annabel.koh@trustsphere.com	1.4000
adesh.goel@trustsphere.com	antony.ebelle@trustsphere.com	4.4000
adesh.goel@trustsphere.com	aravind.mp@trustsphere.com	0.5000
adesh.goel@trustsphere.com	arun.sundar@trustsphere.com	34.3000
adesh.goel@trustsphere.com	bersileus.sacamay@trustsphere.com	0.1000
adesh.goel@trustsphere.com	brian.lebahn@trustsphere.com	4.7000
adesh.goel@trustsphere.com	bryan.acedo@trustsphere.com	0.9000
adesh.goel@trustsphere.com	dawn.radecki@trustsphere.com	3.9000
adesh.goel@trustsphere.com	dev.menon@trustsphere.com	5.7000
adesh.goel@trustsphere.com	gabrielle.tremblay@trustsphere.com	0.1000
adesh.goel@trustsphere.com	gladys.opone@trustsphere.com	0.3000
adesh.goel@trustsphere.com	grace.siew@trustsphere.com	1.7000
adesh.goel@trustsphere.com	graham.wells@trustsphere.com	1.0000
adesh.goel@trustsphere.com	greg.newman@trustsphere.com	3.4000

Standardize and Aggregate to get new trust score

To get our new trust score, we will standardize the feature results and aggregate them to get a single unique result for every pair of employees.

Survey

Mode of data collection: Online survey
Target Sample: All employees in the company (across geographies)
Aim: To measure influence
Summary: The purpose of the survey is to validate the use email exchange network for calculating influence score, where, influence score is defined as the extent to which an individual sways information in the workplace. In a work environment, as there as be different kinds of information flow, we divided the term influence in to six main categories –

1. **Social:** defined as any interaction regarding the business with any colleague. This gives a high level view of the kind of interactions and volumes of interactions between employees.
   → *How many times do you interact with the following colleagues regarding business topics, within a month?*

2. **Information sharing:** defined as an interaction when job related resources or information is transferred between employees.
   → *How many times do you receive job related information from the following colleagues within a month?*

3. **Problem solving:** defined as an interaction where employees seek help in solving problems. These interactions will be dependent on the kind of work-related problems an employee regularly faces.
   → *How many times do you seek help from the following colleagues for business/technical related problems within a week?*

4. **Decision making:** defined as an interaction between two employees where one employee consults the other on a specific business related decision to make.
   → *How many times do you consult the following colleagues if you have a work related decision to make, within a week?*

5. **Support:** defined as an interaction wherein an employee provides career advice to another employee.
   → *How many times do you discuss your career prospects and progression with the following colleagues in a year?*

6. **Idea generation:** defined as an interaction between two employees that involves the discussion of novel ideas or approaches.
   → *How many times do you discuss, share or brainstorm novel ideas with the following colleagues, in a quarter?*

Timeline

Task	Responsible	Wk 9	Wk 10	Wk 11	Wk 12	Wk 13
New Trust Score Creation						
Normalize and standardize features	All					
Correlation analysis of features	All – split features					
Aggregate feature to construct final Trust Score equation	All					
Design Metrics						
Finalize metrics to display on the dashboard on an employee level and relationship level	All					
Hybrid Centrality Algorithm						
Research on different types of centralities	Akshita, Prekshaa					
Run regression analysis for the different centralities	Aayush					
Construct Algorithm	All					
Test against the six aspects of influence from survey	All – split the six aspects					
Finalize Algorithm	All					
Develop Dashboard						
Coding	All					
Deliver to Client + Final						
Write Paper	All					
Report Submission	All					
Presentation Slides	Akshita					