Utilize Communication Data For Network Analysis

May 29, 2024

IMT 542

Chengzi

Agenda

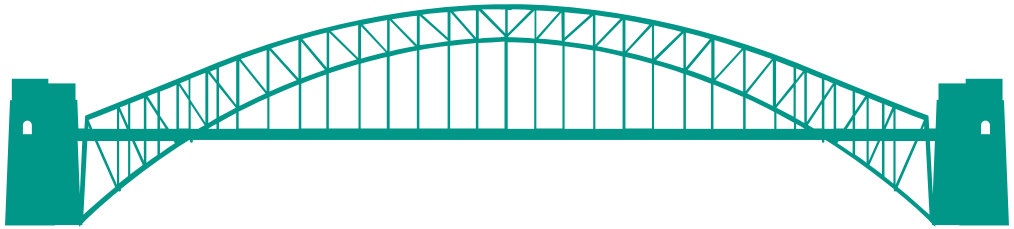
- Introduction about the info story
  - Motivation
- Existing structures
  - Fair assessment
  - Sample input data
  - Sample output data
- Methodology
  - Steps to improve the info structure
- Reflection
  - Takeaways
  - Limitations & future work

Introduction

Goal: To automate the process of cleaning and manipulating data from source to destination, making it ready for network analysis and allowing users to download the data.

Communication Tools

They generate and collect information about a user's communication behaviors.



ONA Tools

They help understand the communication flow through an organization and the interactions among members.

Assessment - FAIR principles

- Findable

  Doing well on F1~4, “Metadata and data are easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services.”

- Accessible

  Doing well on A1: once the user finds the required data, we would know how they can be accessed, possibly including authentication and authorisation.

  Could improve for A2: “Metadata are accessible, even when the data are no longer available.”

- Interoperable

  Doing ok on I1~3: “The data usually need to be integrated with other data, and it needs to interoperate with applications or workflows for analysis, storage, and processing.” Yet, users could not choose the format to access pr download.

- Reusable

  Doing well on R1: “(Meta)data are richly described with a plurality of accurate and relevant attributes.”

Sample Input Data

```
# sample Teams data
teams_data = {
  "teams": [
    {
      "id": "team-id-1",
      "displayName": "Team 1",
      "channels": [
        {
          "id": "channel-id-1",
          "displayName": "General",
          "messages": [
            {
              "id": "message-id-1",
              "createdDateTime": "2023-05-17T10:10:00Z",
              "from": {
                "user": {
                  "id": "user-id-1",
                  "displayName": "John Doe",
                  "userPrincipalName": "john.doe@example.com"
                }
              },
              "body": {
                "content": "Hello, team! This is the first message."
              }
            },
            {
              "id": "message-id-2",
              "createdDateTime": "2023-05-17T10:12:00Z",
              "from": {
                "user": {
                  "id": "user-id-2",
                  "displayName": "Jane Smith",
                  "userPrincipalName": "jane.smith@example.com"
                }
```

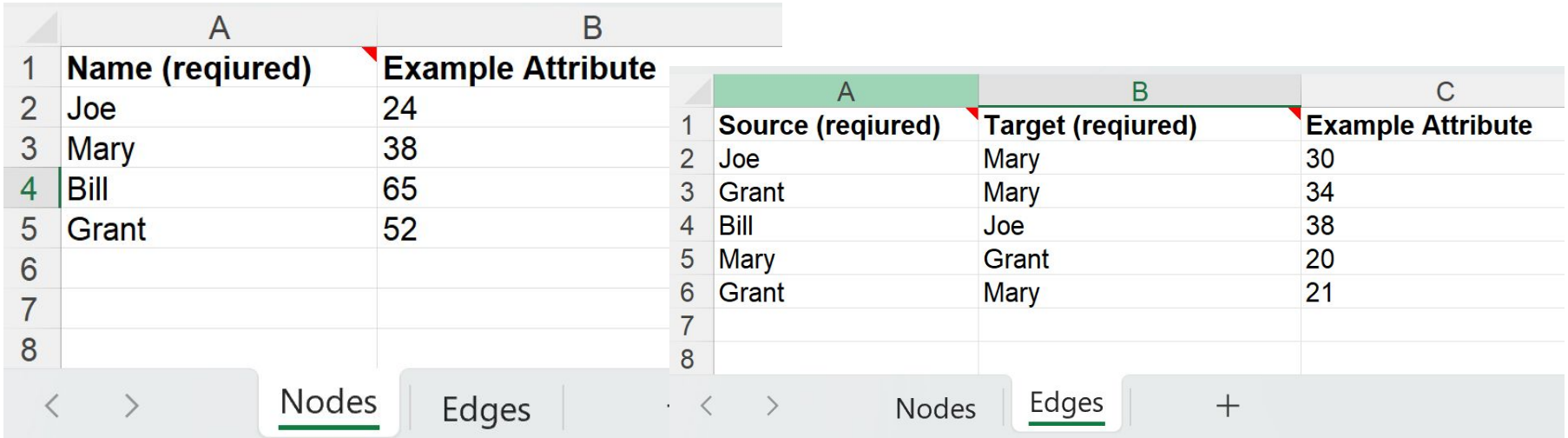
Microsoft Teams (json format)

Gmail (mbox format)

```
X-On-bounce-route-to: notification-service-failures-prod
Date: Thu, 23 May 2024 22:32:24 +0000
From: iSchool Career Services <notifications@instructure.com>
Reply-To:   reply+4a61f32e220a1f3b-10~1267825772-1716503539@notifications.canvasl
To: tiancz23@uw.edu
Message-ID: <0100018fa795817f-b3a64e77-59e4-4e32-860a-88357c6e6326-000000@email.a
Subject: Your Weekly Job Round-Up!: iSchool Career Services
Content-Type: multipart/alternative;
 boundary="---=_mimepart_664fc3f8169f3_68d94a88745df";
 charset=UTF-8
Content-Transfer-Encoding: 7bit
Auto-Submitted: auto-generated
Feedback-ID: ::1.us-east-1.6bnAGrq+H33YQMZ/rl7AdPgJjS+jB1rbcCIER0hXfVM=:AmazonSES
X-SES-Outgoing: 2024.05.23-23.249.221.125
```

Sample Output Data

A Python application that can convert the original (meta)data from Microsoft Teams / Gmail to structured data ready for Polinode.



	A	B
1	**Name (reqiured)**	**Example Attribute**
2	Joe	24
3	Mary	38
4	Bill	65
5	Grant	52
6		
7		
8		

Nodes Edges

	A	B	C
1	**Source (reqiured)**	**Target (reqiured)**	**Example Attribute**
2	Joe	Mary	30
3	Grant	Mary	34
4	Bill	Joe	38
5	Mary	Grant	20
6	Grant	Mary	21
7			
8			

Nodes Edges +

Methodology

Data Extraction

- Use Google Takeout to access the user's account and export email data. *I used this method and gmail mbox data in this project.*

  or

- Use the Export page with the user's Microsoft Account to access and download Teams data.

Data Transformation

- Use Python to load, clean, manipulate, and output data according to Polinode's standards.

  &

- Include documentation to improve process transparency and make it easier for troubleshooting.

Data Transformation

1. Ensure basic data requirements are met before feeding the original data to next steps.
2. Iterate through each message in the Inbox & Sent mbox files, extract key info.
3. Combine datasets.
4. Expand multi recipient edges (since there are cases one person sends message to multiple recipients).
5. Calculate frequency.
6. Clean names to remove email addresses → edges
7. Create a separate data frame with unique email names → nodes
8. Save 2 data frames into one file.

	Date	From	To	Subject
0	Mon, 22 Apr 2024 16:31:58 -0700	Chengzi Tian <tiancz23@uw.edu>	Chengzi Tian <tiancz23@uw.edu>	
1	Sun, 3 Mar 2024 23:58:28 -0800	Chengzi Tian <tiancz23@uw.edu>	tiancz23 <tiancz23@uw.edu>	Draft
2	Mon, 19 Feb 2024 19:44:09 -0800	Chengzi Tian <tiancz23@uw.edu>	hr@itsuggest.com	Application for Business System/Intelligence A...
3	Tue, 23 Apr 2024 17:01:28 -0700	Chengzi Tian <tiancz23@uw.edu>	Chengzi Tian <tiancz23@uw.edu>	resume

Results

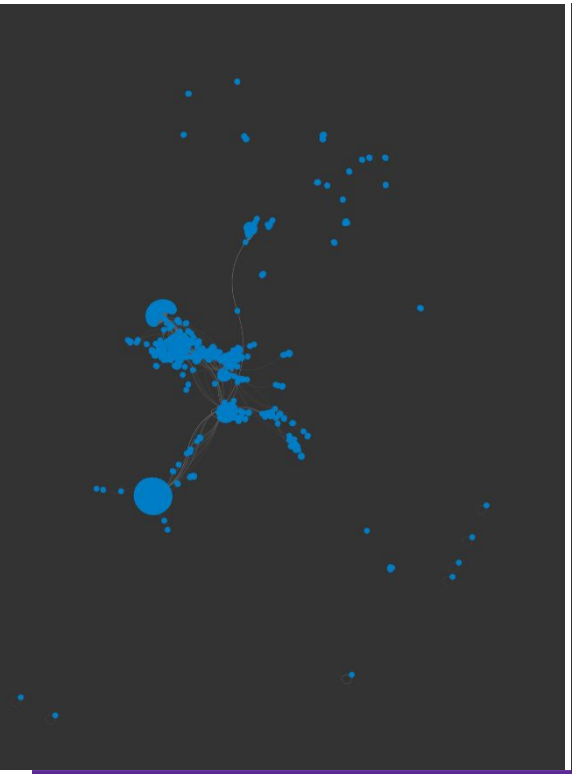
	A
1	Name
2	Adam D Moore
3	alyssa.fatal@bvpri.org
4	Amy Wang
5	Ariella Aung
6	Cameron Cramer
7	Chengzi Tian
8	CHENGZI TIAN personal
9	Chirag Shah
10	Christine Noyes-Williams
11	Ciara Jones
12	Cynthia del Rosario

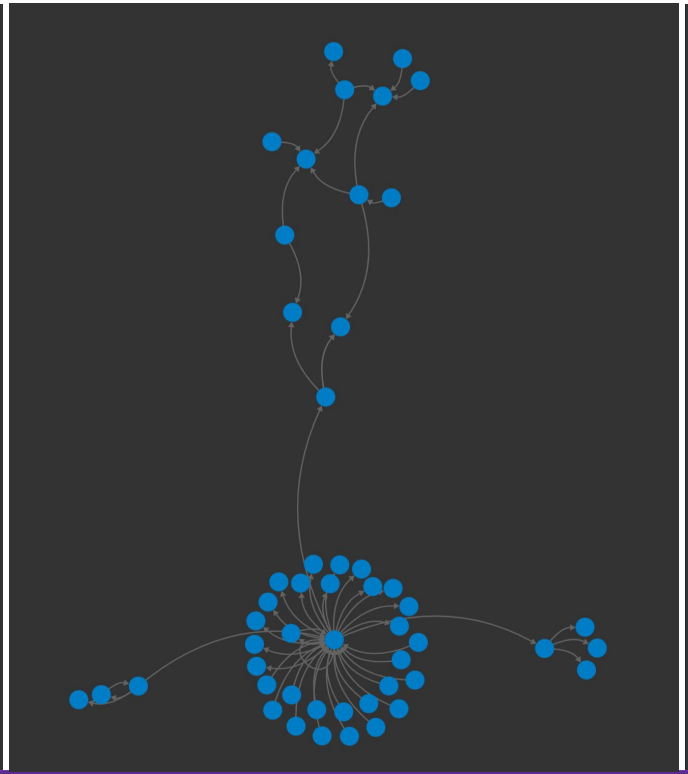
Nodes | Edges

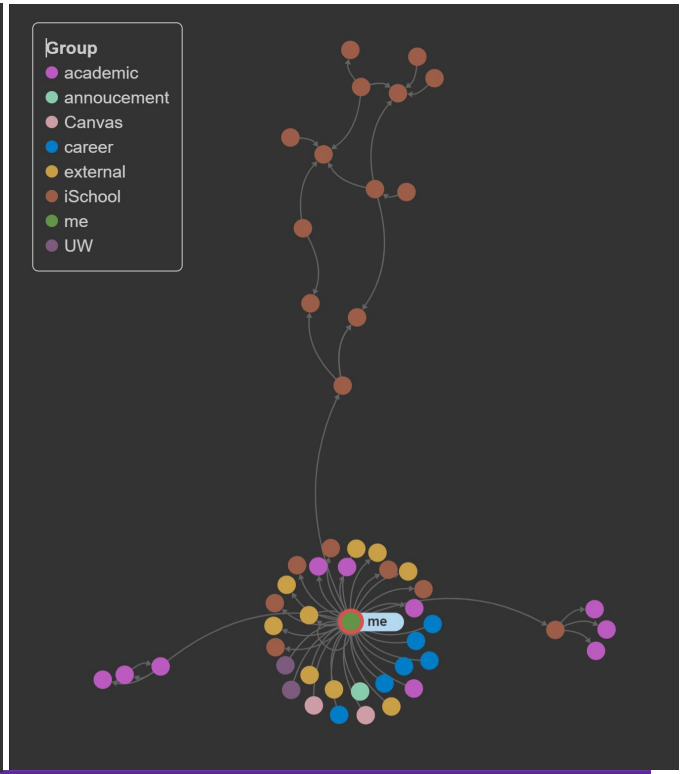
	A	B	C
1	**Source**	**Target**	**frequency**
2	Indeed	Chengzi Tian	111
3	iSchool Career Services	Chengzi Tian	125
4	Your Indeed Job Feed	Chengzi Tian	103
5	Handshake Notification	Chengzi Tian	80
6	Handshake	Chengzi Tian	69
7	Chirag Shah	Chengzi Tian	59
8	Cameron Cramer	imsim@uw.edu	46
9	UW Alert	Chengzi Tian	36
10	Cameron Cramer	imsim@uw.edu	34
11	Cameron Cramer	imsimonline@uw.edu	34
12	mychart.donotreply@uw.edu	Chengzi Tian	28

Nodes | Edges

Results on Polinode







Reflection 1

Impacts & Takeaways:

- Provide a guide for **individual users** to more conveniently access and understand our personal data.
- A great **practice experience** for myself to practice the things I learned from class and become more familiar with different data formats.
- Conquering the last mile could be harder than an initial thought. Thus paying **attention to details** and **testing** are important.

Reflection 2

Limitations & Recommendations for future work:

- **Need a more robust data cleaning process** to address more diversified data contents.
- **More nuanced approach** on different forms of communication such as cc / bcc / conversations and assigning them different weights, so that the visualization would become more outstanding.
- Create **a more user friendly interface,** so that the user could transform their data by clicking their mouse rather than running the program.
- Users could **provide feedback** to the large communication **platforms** and make suggestions based on their user experience, upholding a better environment for portable data.

Thank you!

Reference

- Fair Principles
- Polinode Guide
- Polinode Networks: A Quick Start Guide
- Social Network Analysis Using Enron Email Dataset
- Extracting social networks and contact information from email and the Web
- Export or delete your data in Microsoft Teams
- How to download your Google data
- Gmail Data Analysis using Python
- ChatGPT