Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

`fauci-email`: a `json` Digest of Anthony Fauci's Released Emails

Austin R. Benson,[1] **Nate Veldt,**[2] **David F. Gleich**[3]

[1]Department of Computer Science, Cornell University
[2]Department of Computer Science and Engineering, Texas A&M University
[3]Department of Computer Science, Purdue University
arb@cs.cornell.edu, nveldt@tamu.edu, dgleich@purdue.edu

Abstract

A collection of over 3000 pages of emails sent by Anthony Fauci and his staff were released in an effort to understand the United States government response to the COVID-19 pandemic. In this paper, we describe how the original PDF document of emails was translated into a resource consisting of `json` files that make many future studies easy. We include examples for how to convert this email information into a network, a hypergraph, a temporal sequence, and a tensor for subsequent analysis, and discuss use cases and benefits in analyzing the data in these different derived formats. These resources are broadly useful for future research and pedagogical uses in terms of human and system behavioral interactions.

Introduction

Jason Leopold submitted a freedom of information act request to obtain email surrounding the initial response of United States federal agencies including the National Institutes of Health (NIH) and Centers for Disease Control (CDC) regarding the COVID-19 pandemic. The result was a 3234 page PDF document consisting of emails that Anthony Fauci, the head of the national institute of allergy and infectious disease (NIAID), send between approximately February 2020 and April 2020. The set of released emails immediately became the focus of a large number of subsequent news articles and blog posts, expressing varying opinions and conclusions, often focused on a small subset of email exchanges in the document. However, there has been little effort to translate the dataset into an easier format to process and study, and little attempt to understand the broader structure of the dataset as a whole.

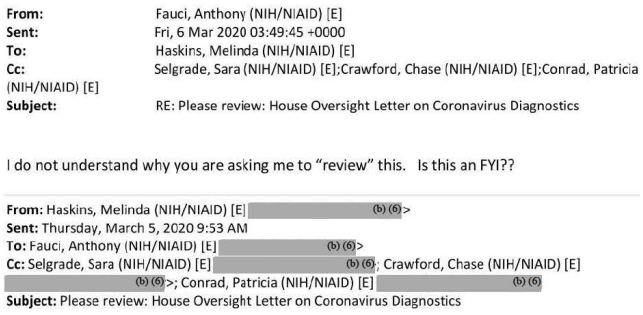
In this paper, we describe how the original PDF was converted into a set of `json` files for subsequent analysis and research. Our data conversion process yields a collection of 1289 email threads with 2761 emails including 101 duplicate emails among the threads. We also present a number of derived datasets, showing how these raw data can be analyzed as a social network or graph, a temporal graph, a hypergraph, or a tensor (a). These processed and freely available datasets include the following:

1. The main `json` digest derived from Bettendorf and Leopold (2021), which has senders and receivers of Fauci's email threads canonically labeled in an easy-to-process format.
2. Five graphs derived from the data from the data ranging from 46 to 869 vertices.
3. A hypergraph derived from the emails themselves with 233 nodes and 254 hyperedges.
4. A temporal sequence of adjacency matrices over 100 days from those 77 people where information can flow among all individuals in a temporally consistent sequence.
5. A tensor projection of the data designed to highlight the role of email carbon copy (CC) networks suitable for hypergraph centrality studies as well as a tensor representation of the data as sender, receiver, time, and word.

In this manuscript, we describe the data conversion process in depth and give an overview of the resulting `json` file. The derived graphs, hypergraph, and tensors that result from these data are small compared with the size of many modern datasets, yet they are not so small as to permit trivial analysis. This renders them a rich setting to investigate what can be ascertained from the data. We discuss different use cases and benefits gained by exploring and analyzing the data in the different formats presented. In an extended technical report (Benson, Veldt, and Gleich 2021), we discuss specific interesting findings in more detail. Here we provide a broad overview of the `json` files and present them as an easy-to-use resources for continued exploration by others in the field.

Similar datasets The most closely related existing dataset is the Enron email dump (Cohen 2004). One of the derived datasets we release is a tensor that can be used in a manner similar to many analyses of the Enron email data. There are also key differences between the two datasets. First, much of the email information in the dataset we release was redacted. Second, we only have *Fauci*'s view on the email instead of raw email inbox dumps for more executives.

There are also many similarities between the derived email network datasets we release and standard benchmark graph datasets used frequently by the network science community. As an example, simple minimum s-t cut analysis used to partition the well-known Karate graph (Zachary


From: Fauci, Anthony (NIH/NIAID) [E]
Sent: Fri, 6 Mar 2020 03:49:45 +0000
To: Haskins, Melinda (NIH/NIAID) [E]
Cc: Selgrade, Sara (NIH/NIAID) [E];Crawford, Chase (NIH/NIAID) [E];Conrad, Patricia (NIH/NIAID) [E]
Subject: RE: Please review: House Oversight Letter on Coronavirus Diagnostics

I do not understand why you are asking me to "review" this. Is this an FYI??

From: Haskins, Melinda (NIH/NIAID) [E] (b) (6)>
Sent: Thursday, March 5, 2020 9:53 AM
To: Fauci, Anthony (NIH/NIAID) [E] (b) (6)>
Cc: Selgrade, Sara (NIH/NIAID) [E] (b) (6); Crawford, Chase (NIH/NIAID) [E] (b) (6)>; Conrad, Patricia (NIH/NIAID) [E] (b) (6)
Subject: Please review: House Oversight Letter on Coronavirus Diagnostics

Figure 1: The first page from the PDF file released as part of the freedom of information act request regarding Fauci's email contains the entirety of Fauci's sent email including information (partially redacted) on the email Fauci was replying to. From this page, we are able to extract information on two emails: (i) an email from Fauci to Haskins with a CC to Selgrade, Crawford, and Conrad on 2020-03-06 and (ii) an email from Haskins to Fauci with a CC to Selgrade, Crawford, and Conrad on 2020-03-05. While we have the text of Fauci's email, the text of the original email is redacted.

1977) can also be used to find almost perfect bipartitions in our derived Fauci email networks (Benson, Veldt, and Gleich 2021). The dataset also shares similarities with the popular Email-EU dataset (Leskovec, Kleinberg, and Faloutsos 2007), which has been studied and analyzed both as a graph (Yin et al. 2017) and as a hypergraph (Benson et al. 2018).

Data Conversion and Processing

Figure 1 contains the first page of the PDF of Dr. Fauci's emails. New emails in the text begin with a *from line* containing "From:", as in Figure 1, to identify the sender of the email. Many email clients include "reply data" in the email information, consequently, we are able to infer some amount of communication outside of only what Fauci sent. For example, the email in Figure 1 shows a reply from Fauci to another group with a long CC list. This is in response to a previous email from the same group.

Conversion and processing The PDF was converted to text and then formatted into a `json` digest. The final digest contains 2,761 emails among 1,309 individuals in 1,289 email threads.

The PDF was first converted to a text file with the `pdftotext` program, specifically, we used the command `pdftotext -layout -r 300 leopold-nih-foia-anthony-fauci-emails.pdf`.

The file was segmented into chunks of text corresponding to email threads; the start of a thread was considered to be a "from" line with Fauci as sender that also began with a form feed character (indicating a new page of the pdf). The emails within a thread were found by "from" lines.

The start of the emails contained clear delimiters for the sender, timestamp, recipient list, cc list, and subject (Figure 1). The body of the email was then taken to be all text after the subject and before the next email in the thread.

Timestamps appeared in ten different formats that could be parsed by Python's `datetime.strptime` function. The main challenge was handling the numerous errors in the PDF to text conversion. For example, "Thursday" might appear as "Thu rsday" or the number 1 and letter l were often interchanged. Parsing the timestamp involved several general string substitutions and many manual rules for special cases. We successfully parsed timestamps for 86.5% of identified emails, and we omitted emails for which we could not parse a timestamp.

The sender, recipient list, and cc list were handled similarly. For the recipient and cc lists, individuals were separated by the semicolon ';' (the cc list in Figure 1 has two semicolons for the three individuals). Standardizing names involved both automation and considerable manual inspection. There were issues with text conversion; for instance, "fauci" was parsed into several textual variants, including "f auci," "f.aucl," "fa uci," "fa11ci," and "fauc i." Also, one individual could appear with multiple variants on their name or address. For example, the individual Cliff Lane appeared as "Lane, Cliff," "Cliff Lane," and "clane@niaid.nih.gov" in different emails. The standardization process was iterative. Given a tentative list of names, we used matching algorithms to find possible duplicates, and these were often checked by manually inspecting the PDF. Sometimes, emails were sent on behalf of someone else (e.g., Patricia Conrad on behalf of Anthony Fauci). We treated these as their own "names" rather than attributing to one of the parties. We omitted any emails where we could not identify a sender or at least one recipient, which occurred in 5.1% of the cases. The omissions were mostly caused by redactions or severe errors in the PDF to text conversion.

We also identified federal organizations to which individuals belonged via designations in the email names (e.g., "NIH" appearing after all names in Figure 1). Organization affiliations were National Institutes of Health (NIH), Health and Human Services (HHS), Centers for Disease Control and Prevention (CDC), the Food and Drug Administration (FDA), Office of the Secretary (OS), and the Executive Office of the President (EOP). Around 26.6% of individuals were identified as belonging to one of these organizations, and all of the memberships were manually verified. Many of these organizations are in fact contained within one another; for example, the CDC, NIH, and FDA are all part of the HHS. When associating individuals to organizations, we simply identified the first organizational affiliation listed directly next to the individual in the email header. For example, Anthony Fauci appears as "Fauci, Anthony (NIH/NIAID)", so is assigned to the NIH cluster.

Additional manual processing and challenges As highlighted above, our processing of dates, name standardization, and identification of organization affiliations all involved a significant amount of iterative semi-manual processing and manual verification. Our automated processing also produced a number of cases where the subject line was stored incorrectly, usually due to difficulties parsing

redacted information in the sender, recipient, and CC list information. For example, in some cases an email account in the CC list would be parsed as the subject, and the subject itself would be parsed as part of the email body. We performed a careful manual check to fix these issues. We first printed out and individually inspected all nonempty subject lines that did not begin with "Re", "RE", "Fw", or "FW". Of these, we flagged 25 subject lines that appeared potentially erroneous (not counting simple OCR errors), and cross checked the automatically parsed data for these emails against the original pdf. Of these, there were 5 examples where the CC list was incorrectly parsed as empty, while the real CC recipient was moved to the subject and the subject was moved to the email body. There were 2 examples where part of the "To:" field was incorrectly parsed as part of the subject. There were also 3 examples where the subject was moved to the email body but no errors were made in sender, receiver, or CC list information. There were also 57 emails in the original automated parsing that were empty. We cross checked each of these manually against the original pdf, and found 29 cases where the subject was moved to the email body but no other errors were made, 16 emails that has some type of mistake in the sender/receiver/cc information, and 12 cases where there was no error. In total, there were 23 emails with subject line errors that included an error in sender/receiver/cc information, and a handful of others that included smaller errors. This constitutes only 0.833% of the 2761 emails in the dataset. We manually fixed all of these errors for the final version of the dataset we released.

We also noted errors in parsing some timestamps, including a number of cases where an email sent with a PM timestamp was parsed as having been sent in the AM. Even without these parsing errors, timestamps are challenging to properly record as they come from various different time zones. The email body text still contains OCR errors, and one unavoidable challenge in parsing the data is that a significant portion of the PDF text is redacted. Independent parsing strategies may lead to slightly different counts for emails and threads. As is the case for many similar datasets, such as the popular Enron email dump (Cohen 2004), there may be future releases of this data that help correct errors further.

FAIR Principles and Ethical Considerations Our datasets abide by FAIR principles: they are findable, accessible, interoperable, and reusable. *Findable and Accessible.* Our datasets are published on the Zenodo website and freely accessible at https://doi.org/10.5281/zenodo.5828209. The datasets and code for processing them are also publicly available on a public GitHub repository https://github.com/nveldt/fauci-email. *Interoperable and Reusable.* We purposely chose the familiar and widely-used `json` file format for ease of use by other researchers. The provenance of our data is the original PDF of emails that is already publicly available online. Furthermore, in our public GitHub repository we include all code needed to convert the original PDF into our derived `json` files, including detailed explanations and documentation of the manual fixes involved. Our entire data processing pipeline can be viewed, checked, and reused by other researchers, who may also process and store the data in alternative ways if desired.

Description of the Fauci-Email `json`

The Fauci-email `json` dataset stores the following four key-value pairs:

- `names`: An array of names for email accounts involved in the dataset (e.g., "fauci, anthony", "condrad, patricia", or "mmwrmedia@listserv.cdc.gov").
- `clusters`: An array of organization affiliation labels for accounts in the `names` array (e.g., the label for "fauci, anthony" is 1, the label corresponding to the NIH).
- `cluster_names`: An array with 7 entries providing the name for each organization affiliation: `["NIH", "HHS", "CDC", "FDA", "OS", "EOP", "other"]`.
- `emails`: an array of email threads.

The ith entry of `emails` is itself an array of email objects, corresponding to the ith email thread from the PDF. Each individual email object is associated with the key/value pairs summarized in Table 1. A thread in the `json` file has more than one email if our parsing detected multiple "from" lines between the start of different threads, each corresponding to individual emails within the thread. For example, the first thread in our file corresponds to the thread in Figure 1 involving two emails: the email from Melinda Haskins with subject line "Please review: House Oversight Letter on Coronavirus Diagonostics", and Dr. Fauci's reply to this email. Another thread with multiple emails is shown in Figure 3.

Basic dataset statistics. Table 2 displays the top 10 senders, recipients, and cc-recipients in the dataset. While many individuals are ranked highly in all three lists, there are also many others who appear frequently in only one of these roles. For example, *niaid odam* is a mailing list that is frequently forwarded emails for discussion, but is never a sender and is on the cc list much less often than the primary recipient list. There are also several NIH employees who are frequently CCed on emails (e.g., Hilary Marston, Kathy Stover, Kimberly Barasch) but rarely act as senders or primary recipients.

Figure 2 shows a plot of the number of emails in the dataset per day, based on collected timestamps. Note that these are counts for the number of emails in the released PDF for which we were able to obtain a timestamp. Dr. Fauci likely received and sent many other emails that are not included in the PDF, and therefore are not in our `json` dataset.

Summary of key people. We provide a briefly annotated list of key individuals to help contextualize dataset statistics and results.

Anthony Fauci Head of the National Institute of Allergy and Infectious Disease (NIAID), a group within the US National Institutes of Health (NIH).

Field	Description	Example
recipients	Name IDs for recipients	`1`
body	Email body text	`I do not understand why you are asking me to "review"...`
timestamp	Email timestamp	`2020-03-06T03:49:45+00:00`
sender	Name ID for sender	`0`
cc	Name IDs for CCed accounts	`[2, 3, 4]`
subject	Email subject text	`RE: Please review: House Oversight Letter on Coronavirus...`

Table 1: Summary of data entries for each individual email, and example for the first email in the PDF (see Figure 1). Recipients, sender, and cc-recipients IDs are indexed from zero, i.e., `names[0] = "fauci, anthony"`.

	Sender		
1	1	8	fauci, anthony, 1287
2	2	1	conrad, patricia, 81
3	3	39	collins, francis, 49
4	5	4	billet, courtney, 45
5	80	285	mecher, carter, 36
6	6	2	folkers, greg, 28
7	–	–	eisinger on behalf of fauci 22
8	11	9	tabak, lawrence, 18
9	14	7	routh, jennifer, 18
10	30	127	goldner, shannah, 16

	Receiver		
1	1	8	fauci, anthony, 982
2	2	1	conrad, patricia, 277
3	3	39	collins, francis, 129
32	4	46	cassetti, cristina, 122
4	5	4	billet, courtney, 84
6	6	2	folkers, greg, 79
25	7	13	lerner, andrea, 61
–	8	62	niaid odam, 58
81	9	61	auchincloss, hugh, 53
19	10	11	lane, cliff, 52

	CC		
2	2	1	conrad, patricia, 366
6	6	2	folkers, greg, 146
18	12	3	marston, hilary, 99
4	5	4	billet, courtney, 98
30	24	5	stover, kathy, 87
208	22	6	barasch, kimberly, 85
9	14	7	routh, jennifer, 78
1	1	8	fauci, anthony, 62
8	11	9	tabak, lawrence, 47
219	17	10	eisinger, robert, 46

Table 2: Top ten senders, recipients, and cc-recipients in the `json` dataset. The first three columns in each table list the rank of each email account as a sender, recipient, and cc-recipient respectively. A dash indicates no participation in this role. The number of times the account participated in this role is listed by the name of the account or individual.

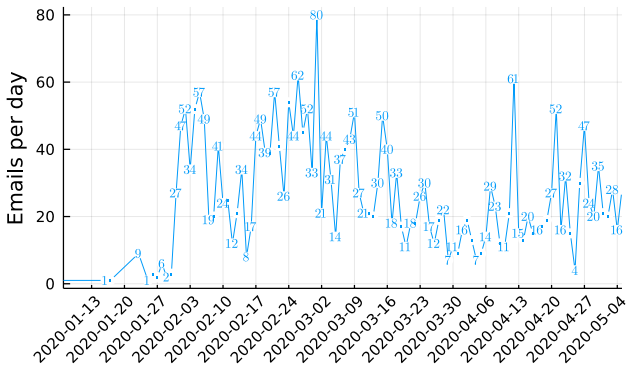


Figure 2: Number of emails per day in the Fauci email `json` dataset.

Patricia Conrad Fauci's key special assistant and frequent email proxy.

Francis Collins Head of the US National Institutes of Health (NIH), the NIH are an organizational division of the US Department of Health and Human Services (HHS).

Robert Redfield Head of the US Centers for Disease Control (CDC), the CDC are another organizational division of HHS.

Alex Azar Secretary (head) of the US Department of Health and Human Services (HHS), part of the president's cabinet.

Robert Kadlec Assistant Secretary of Preparedness and Response for HHS.

Deborah Birx The Coronavirus Response Coordinator appointed by the US President and a member of the Coronavirus Task Force.

Jennifer Routh Science communication editor in the NIAID division of the NIH.

Greg Folkers Anthony Fauci's chief of staff.

Lane, Cliff a clinical director at NIAID.

Billet, Courtney was often CCed as a point of coordination for Fauci replying to reporters.

Derived Datasets and Use Cases

The original `json` file of email data can be transformed into different types of derived datasets that are suitable for many types of studies at the intersection of sociology and network science. We describe several examples, with illustrations of use cases. All of these derived datasets are available with the master `json` file on Zenodo (https://doi.org/10.5281/zenodo.5828209).

Network Datasets

The data can be modeled in terms of a number of different networks that we describe here. Note that there are many other possible networks. For instance, although Fauci was removed from many of these networks, they all could have Fauci in them too. Many of these networks can be viewed as

projections of some hypergraph. A hypergraph is a generalization of a graph in which arbitrary-sized subsets of nodes can be joined in a multiway relationship called a *hyperedge*. A hyperedge involving only two nodes corresponds to the standard notion of an edge in a graph. Projection techniques are designed to reduce a hypergraph to a related graph by replacing hyperedges with weighted edges.

`repliedto-nofauci` This is a weighted network that enumerates *replied-to* relationships. We have an edge from u to v if u replied to v's email and then weight the edge with the largest number of interactions in either direction. We remove Fauci from this view of the network to study the view without his emails. This network is an instance of a temporal motif network using a "replied-to" temporal motif (Paranjape, Benson, and Leskovec 2017). We then remove everyone outside of the largest connected component.

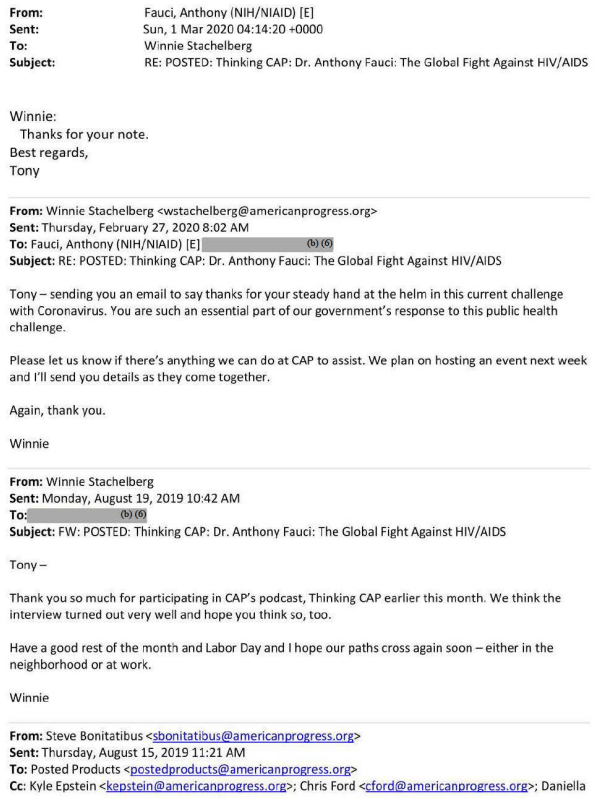
`tofrom-nofauci-nocc` This is a weighted network that has an edge between the sender and recipients of an email (excluding the CC list), weighted by the largest number of interactions in either direction. In this network, we remove emails with more than 5 recipients to focus on *work* behavior instead of *broadcast* behavior. This omits, for instance, weekly emails that detail spending of newly allocated funds to address the pandemic that were often sent to around 20 individuals. We also remove everyone outside the largest connected component.

`tofrom-nofauci` This is the same network above, but expanded to include the CC lists in the number of recipients. The same limit of 5 recipients applies.

`hypergraph-projection-nocc` This is a weighted network that is a network projection of the email hypergraph where each email indicates a hyperedge among the sender and recipients. We then form the clique projection of the hypergraph, which means that each hyperedge is replaced by a fully connected set of edges among all participants to form a graph. The weight on an edge in the network are the number of hyperedges that share that edge. The graph is naturally undirected. Because this omits CC lists from each hyperedge, the graph can easily be disconnected if an email arrived via a CC edge. To focus the data analysis, we remove any individual who has only a single edge in the graph (with any weight).

`hypergraph-projection` This version of the network adds CCed recipients to the hyperedge for each email. This remains disconnected largely due to email lists and BCC-events in the data (see Figure 3 for an instance of a list on page 128 and page 1508 in the PDF Bettendorf and Leopold (2021) for an instance of a BCC) even though Fauci remains in this data. Other disconnections are due to parsing errors. There are 35 nodes that are removed due to disconnection.

These are all weighted networks. Consequently, they can be analyzed as both simple networks (with edge weights and self-loops removed) or weighted networks depending on the type of analysis. Basic statistics of the networks are given in Table 3.



From: Fauci, Anthony (NIH/NIAID) [E]
Sent: Sun, 1 Mar 2020 04:14:20 +0000
To: Winnie Stachelberg
Subject: RE: POSTED: Thinking CAP: Dr. Anthony Fauci: The Global Fight Against HIV/AIDS

Winnie:
Thanks for your note.
Best regards,
Tony

From: Winnie Stachelberg <wstachelberg@americanprogress.org>
Sent: Thursday, February 27, 2020 8:02 AM
To: Fauci, Anthony (NIH/NIAID) [E] (b) (6)
Subject: RE: POSTED: Thinking CAP: Dr. Anthony Fauci: The Global Fight Against HIV/AIDS

Tony – sending you an email to say thanks for your steady hand at the helm in this current challenge with Coronavirus. You are such an essential part of our government's response to this public health challenge.

Please let us know if there's anything we can do at CAP to assist. We plan on hosting an event next week and I'll send you details as they come together.

Again, thank you.

Winnie

From: Winnie Stachelberg
Sent: Monday, August 19, 2019 10:42 AM
To: (b) (6)
Subject: FW: POSTED: Thinking CAP: Dr. Anthony Fauci: The Global Fight Against HIV/AIDS

Tony –

Thank you so much for participating in CAP's podcast, Thinking CAP earlier this month. We think the interview turned out very well and hope you think so, too.

Have a good rest of the month and Labor Day and I hope our paths cross again soon – either in the neighborhood or at work.

Winnie

From: Steve Bonitatibus <sbonitatibus@americanprogress.org>
Sent: Thursday, August 15, 2019 11:21 AM
To: Posted Products <postedproducts@americanprogress.org>
Cc: Kyle Epstein <kepstein@americanprogress.org>; Chris Ford <cford@americanprogress.org>; Daniella

Figure 3: An example email exchange that produces a disconnected component. In this case, a mailing list "posted products" generated an email to multiple people, that were forwarded to Fauci. But Fauci is disconnected from the original email. This could be addressed by adding links based on the threading, although we did not pursue this avenue in our analysis.

Example use case: cluster analysis Representing the Fauci email dataset as a network allows us to apply standard network analysis tools to analyze the structure of email interactions. We illustrate this by computing centrality scores and performing cluster analysis on the simple graph `tofrom-nofauci-nocc`. We find that the clusters formed by solving the modularity graph clustering objective to optimality (Newman and Girvan 2004) are characterized by nodes of high betweenness centrality (Freeman 1977; Csardi and Nepusz 2006) that identify functions and groups in the emails. See Figure 4, where we label nodes with high betweenness centrality.

This analysis shows that agency heads (Collins, Redfield) and task coordinators (Birx, Farrar) are high betweenness nodes in distinct clusters. The clusters identified revolve around different agencies (NIH, CDC, WHO) or functional tasks (handling media requests, budgets), or involve email exchanges around a specific topic, for instance an editorial for the New England Journal of Medicine. Remember that Fauci is involved in almost all of the emails, so the interaction between Redfield, Collins, and Farrar is really modu-

graph	nodes	simple graph					weighted graph						
		edges	max deg	mean deg	med deg	λ_2	loops	vol	loop vol	max wdeg	mean wdeg	med wdeg	λ_2
`repliedto-nofauci hypergraph...`	46	58	18	2.5	1	0.0167	2	435	7	91	9.5	3	0.0082
`-proj-nocc`	372	2589	267	13.9	6	0.0536	0	13120	0	1998	35.3	11	0.0346
`-proj tofrom...`	891	7250	697	16.3	7	0.0084	0	76910	0	4524	86.3	11	0.005
`-nofauci-nocc`	233	325	44	2.8	1	0.0331	2	1168	2	102	5.0	2	0.0309
`-nofauci`	386	585	97	3.0	2	0.0438	9	2173	15	247	5.6	2	0.0316

Table 3: The 5 canonical graphs we derive from the email data along with some simple statistics. Each graph is connected, and there is a simple version without weights and self-loops along with a weighted version that has integer edge weights along with possible self-loops. The number of edges is the count of undirected edges, so there are twice this many non-zeros in the adjacency matrix of the simple graph. The weighted graph also has loops, which gives twice this many non-zeros plus the number of loops in the adjacency matrix. We also show the total volume (sum of weighted degrees) of the weighted graph along with max, median (med), and mean statistics on the degrees of the simple (deg) and weighted graphs (wdeg). Finally, we show the value of λ_2 associated with the normalized Laplacian matrix. The graph names with `nofauci` do not include Fauci's node and those with `nocc` omit the CC lists from the construction whereas those without this treat *CC* lists equivalently with other recipients.

lated by Fauci as well, despite the appearance in this network otherwise. Overall, this shows the power of this type of analysis to identify relevant structure in these networks with only a little information. In these networks, the optimal modularity partitions feature nodes with large betweenness centrality, showing how this network appears to be constructed with local leaders as one might expect in a working hierarchy.

Our extended technical manuscript (Benson, Veldt, and Gleich 2021) illustrates other analyses that can be performed using network representations of the data. This includes a comparison of PageRank centrality scores in different networks, nearly balanced splits in many derived graphs that result from finding a minimum graph cut that separates Francis Collins (head of the NIH) from Patricia Conrad (Fauci's key assistant), and interesting differences between partitions obtained by optimally solving two closely related graph partitioning objectives: *normalized cut* and *conductance*.

Hypergraph Dataset

The `hypergraph-projection` data is one example of a hypergraph analysis (as a projected graph). We additionally provide an explicit way to model the dataset as a hypergraph where each email is a hyperedge among the senders and recipients (excluding the CC entries) – excluding Fauci. We remove any individual that does not have at at least degree 5 in a clique expansion of the hypergraph. The largest connected component of the resulting hypergraph has 233 vertices and 254 hyperedges.

Example use case: differences between local diffusions A local diffusion in a graph or hypergraph answers the question: *what else might be related to a given node in a graph or hypergraph*. It's an instance of a relationship-by-transitivity-of-relationships study. Local diffusion analysis on hypergraphs have been a recently active area. Here, we show how three closely related ideas around PageRank-like diffu-

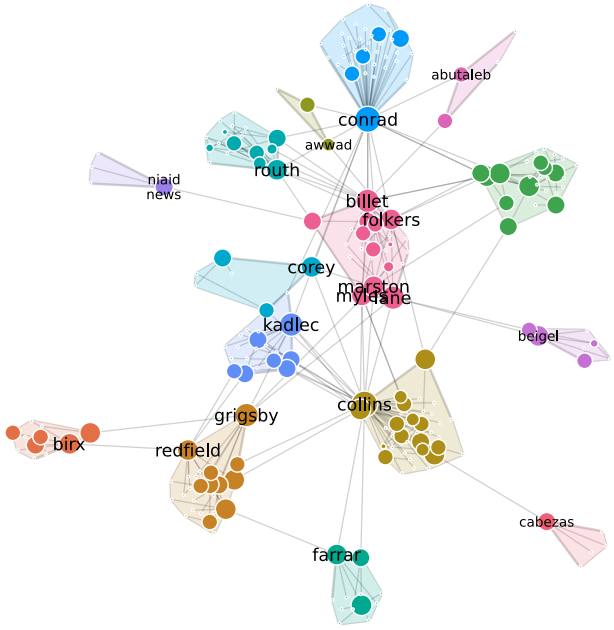


Figure 4: The optimal modularity partition of the network of senders and receivers alone (without Fauci) and reduced to a simple graph are indicated by the colored regions. There are 15 groups and the layout is designed to highlight the modularity groups. We show the 14 most central nodes by betweenness centrality scores in a large fontsize, which labels at least one vertex in all but 5 groups. The small fontsize labels on Abutaleb (rank 46), Awwad (rank 76), Beigel (rank 24), Cabezas (rank 28), and niaid news (rank 33) show key nodes in clusters that were not top 14 betweenness. Note that many of the agency heads and task leads are identified as key nodes in these networks (Collins, Redfield, Birx, Farrar).

sions produce strikingly different results on this hypergraph, which indicates it's a useful tool for followup work on comparisons among the implications of these ideas.

PageRank-like diffusions are *quadratic* or *smoothed* variations on cut problems for graphs and hypergraphs (Liu et al. 2021). They can be *seeded* on a single node to generate a ranked list of other nodes based on relationship strength. We do this for a sparse PageRank diffusion on a graph projection of the hypergraph, a direct sparse PageRank diffusion on the hypergraph, and a unregularized PageRank diffusion on the hypergraph. (Sparse PageRank diffusions include regularization extra terms to encourage sparse solutions of the PageRank diffusion equations.) The difference in results is shown in Table 4. There are far more differences than one would expect between these solutions. This indicates an area of further study, and shows that this dataset can provide an interesting case study for exploring how and why related diffusion techniques produce very different results in practice. It is possible that simple parameter changes or other tools will show how these are more similar than apparent from this simple experiment.

Temporal Graph Dataset

We processed the data in a set of directed edges for emails that were sent on the same day, restricted to the largest temporal strong component. A temporal strong component (Bhadra and Ferreira 2003; Nicosia et al. 2012) of a temporal graph is a set of nodes where there is a time-respecting path among all vertices in the component. For the Fauci email dataset this gives a set of 77 nodes. For these nodes, we obtained a sequence of 100 adjacency matrices for each day from February 1 2020 to May 5 2020 with a few other preliminary days (e.g. a September 4, 2018 email from Folkers to Fauci on CDC guidelines on aerosol protections for influenza and coronaviruses, Page 429).

Example use cases: temporal communicability and temporal modularity The first analysis we did was a temporal communicability analysis (Grindrod et al. 2011). This analysis scores each node based on a weighted average of the length of email chains they start (broadcast-centrality) or receive (receive-centrality). The results in Table 5 highlight differences between the broadcasting and receiving role each individual plays. For example, we see that Cristina Cassetti, a program officer at the NIH, has a much higher "receive" role than a "broadcast" role. Searching the dataset for emails involving Cassetti reveals a large number of emails in which Dr. Fauci directly forwards an email to Cassetti, often simply with the message "Please handle". Meanwhile, Jeremy Farrar, the director of a charitable health research foundation named Wellcome Trust, has a much higher role as broadcaster than receiver. Farrar was previously identified as a high betweenness centrality node in our community analysis experiments (Figure 4).

The second analysis was a temporal community analysis (Mucha et al. 2010). This analysis assigns a community or group to each node at each time-point to reflect how the groups change over time. Formally, this is a modularity-like analysis on a temporally-linked graph – this allows the analysis to violate a strict arrow of time and foreshadow the future. The communities this analysis identifies show how the emails sent respond to various external events (Figure 5).

We also created a force directed animation of this dataset to illustrate the temporal modularity groups. This animation is available from our GitHub repository https://github.com/nveldt/fauci-email/blob/master/figures/anim-mod.mp4.

Tensor Datasets

Finally, we present a derived tensor dataset which facilitates exploration of the higher-order structure in the emails through sender–receiver–CC interactions. We first found a maximal set of nodes where everyone participates in the sender, receiver, and CC roles with all of the other nodes in the set. Specifically, we examine all emails containing at least one recipient and at least one CC and find the set of discard nodes D corresponding to people that do not appear at least once as a sender, receiver, and CC in these emails. After, we discard emails where a node in D is a sender, and omit nodes in D from the recipient and CC lists of the other emails. This process is repeated until there are no nodes in the discard set. In the end, there remained a set S of 44 nodes and 1,413 emails with a sender, at least one recipient, and at least one CC from S.

We next constructed a $44\times44\times44$ (non-symmetric) tensor T representing the email relationships of the nodes S. Let s_i represent the sender of the ith email and r_i and c_i the subsets of S who are recipients and CC. Then the tensor entries map the total email volume of the nodes, scaled by the number of email participants:

$$T_{u,v,w} = \sum_i \frac{1}{|c_i| \cdot |r_i|} I(u \in c_i) I(v \in r_i) I(w = s_i),$$

where $I(\cdot)$ is the indicator function.

We also release two tensors that mirror many analyses of the Enron email data (Cohen 2004) where we examine interactions among sender, receivers, time, and words. We first compute word embeddings for 627 commonly used words in the email corpus using word2vec (Mikolov et al. 2013). We create a $1309 \times 1309 \times 102 \times 627$ tensor (`fauci-email-tensor-words.json`) where entry (i, j, k, l) has value v if person i sends an email to person j, k days after date "2020-1-15", and the embedding of the email body is most similar to word l. The value v is the inverse of the number of recipients in the email. We also release a smaller subtensor of size $77 \times 77 \times 96 \times 337$ (`fauci-email-tensor-words-tsc.json`), which excludes emails in the first week (during which time not many emails were sent), and only considers emails among members of the largest temporal strong component and the subset of words that appear in this case.

Example use case: tensor centrality scores We computed the hypergraph H-eigenvector centrality scores (Benson 2019) for the $44 \times 44 \times 44$ tensor T, which is a positive unit-1-norm (unit-sum) vector x such that

$$\lambda x_u^2 = \sum_{v,w} T_{u,v,w} x_v x_w$$

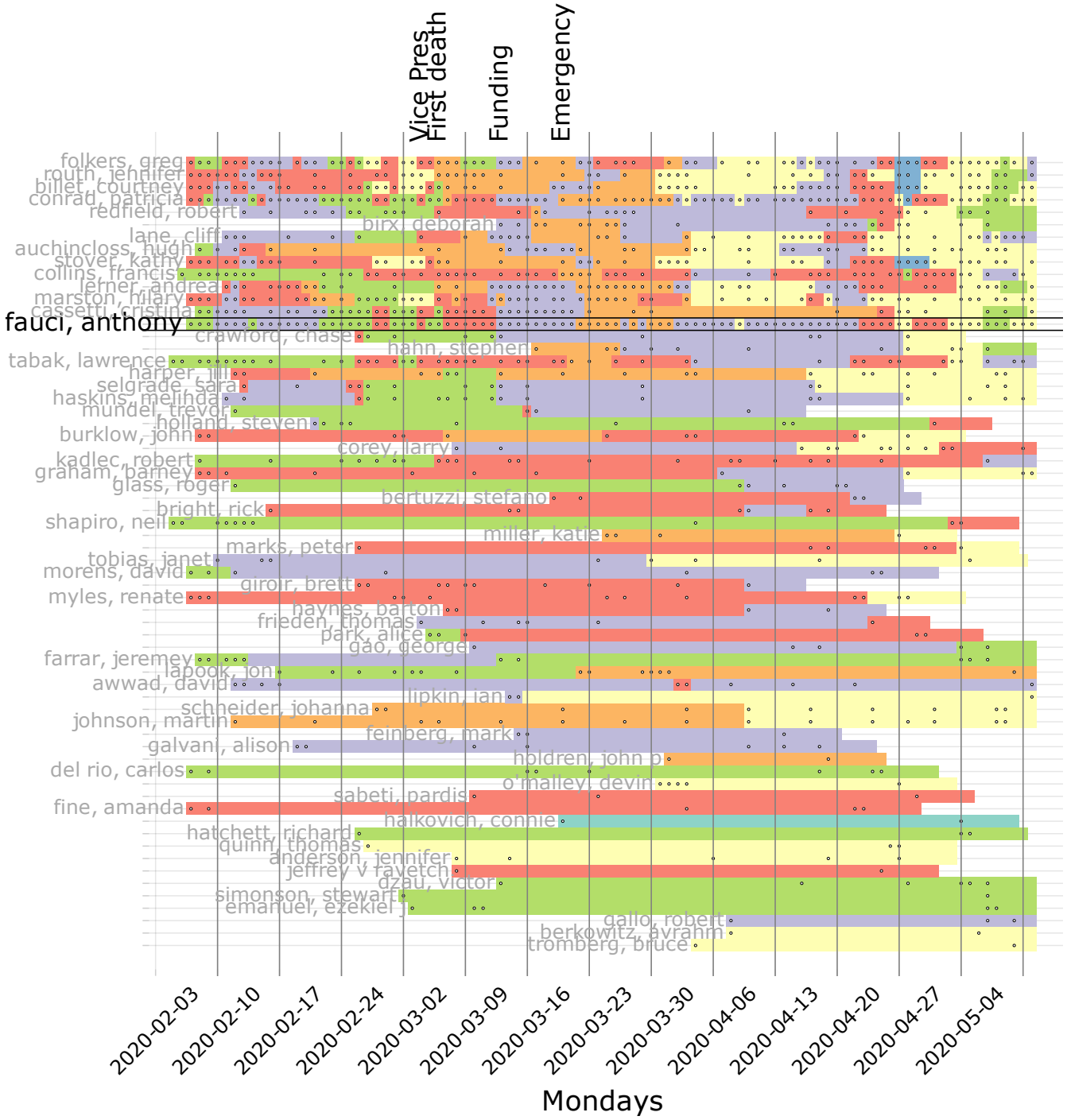


Figure 5: A plot of the communities in a temporal modularity analysis of the network; the figure should be viewed zoomed in and studied for best effect. There are 7 groups, indicated by colors. Nodes are sorted by the number of distinct communities they are a part of, so the first few nodes switch between communities through the time-course of the emails. Community assignments are hidden until the node sends their first email and the small circles indicate days the individuals sent email along with 7 days after their last email. A few key dates are listed at top. The "Vice Pres" event is when Vice President Pence was appointed head of the Coronavirus Task Force; the first death of an American with COVID-19 was on Feb 28; there was a supplemental funding package passed on March 6, 2020; and there was a national emergency declaration on March 13, 2020. Fauci's node is highlighted in the middle.

Sparse Seeded PageRank-Graph				Sparse Seeded PageRank-HyperGraph				Seeded PageRank-HyperGraph			
1	1	1	conrad, patricia, 0.150576	1	1	1	conrad, patricia, 0.415191	1	1	1	conrad, patricia, 0.200878
2	26	3	billet, courtney, 0.031066	50	2	18	hynds, joanna, 0.319236	3	19	2	folkers, greg, 0.061595
3	19	2	folkers, greg, 0.023742	6	3	6	goldner, shannah, 0.319236	2	26	3	billet, courtney, 0.050811
4	145	4	collins, francis, 0.019172	98	4	34	koerber, ashley, 0.319236	4	145	4	collins, francis, 0.05054
5	83	8	lane, cliff, 0.018066	66	5	23	katz, ruth, 0.319236	16	20	5	niaid odam, 0.039361
6	3	6	goldner, shannah, 0.016818	29	6	12	figliola, mike, 0.312654	6	3	6	goldner, shannah, 0.038131
7	24	115	brennan, patrick, 0.016648	139	7	20	barasch, kimberly, 0.216288	18	69	7	auchincloss, hugh, 0.025701
8	37	10	marston, hilary, 0.016041	128	8	101	rom, colin, 0.180251	5	83	8	lane, cliff, 0.024322
9	21	27	lepore, loretta, 0.015216	129	9	100	amerau, colin c, 0.180238	10	34	9	routh, jennifer, 0.021993
10	34	9	routh, jennifer, 0.014553	130	10	99	gathers, shirley, 0.180224	8	37	10	marston, hilary, 0.0205
11	18	47	bonds, michelle, 0.014437	125	11	102	good-cohn, meredith, 0.18021	21	144	11	tabak, lawrence, 0.018368
12	23	121	fine, amanda, 0.014376	126	12	103	mcguffee, tyler ann, 0.180195	29	6	12	figliola, mike, 0.015217
13	95	17	kadlec, robert, 0.01373	127	13	104	edwards, sara l, 0.180179	20	77	13	erbelding, emily, 0.014603
14	164	21	redfield, robert, 0.012563	131	14	74	deatrick, elizabeth, 0.164756	15	39	14	awwad, david, 0.012806
15	39	14	awwad, david, 0.011276	95	15	32	harris, kara, 0.140996	28	73	15	niaid ocgr leg, 0.011777

Table 4: Seeded PageRank and Sparse PageRank results on a graph (left) and hypergraph (middle and right) show surprising differences among the highly ranked nodes of the diffusion – indicating this is a useful dataset for further study. We do this for a sparse PageRank diffusion on a graph projection of the hypergraph, a direct sparse PageRank diffusion on the hypergraph, and a unregularized PageRank diffusion on the hypergraph, all seeded on Patricia Conrad.

broadcast			receive		
1	3	fauci, anthony, 208.8	2	1	conrad, patricia, 132.0
2	1	conrad, patricia, 58.9	7	2	folkers, greg, 61.5
3	4	billet, courtney, 50.9	1	3	fauci, anthony, 60.6
4	29	farrar, jeremey, 47.8	3	4	billet, courtney, 45.0
5	7	collins, francis, 42.1	20	5	lerner, andrea, 30.6
6	10	routh, jennifer, 27.0	19	6	lane, cliff, 29.7
7	2	folkers, greg, 23.6	5	7	collins, francis, 29.5
8	12	tabak, lawrence, 16.3	24	8	cassetti, cristina, 28.9
9	21	myles, renate, 15.2	16	9	marston, hilary, 27.7
10	24	lapook, jon, 13.4	6	10	routh, jennifer, 26.6

Table 5: Among the 77 nodes in the largest temporal strong component, the top 10 nodes by temporal sender and receiver centrality (Grindrod et al. 2011) with parameter 0.02 show Fauci and Conrad as the top broadcast and receiver nodes, respectively. The light fontcolor indicates the rank in the sorted list and the dark fontcolor indicates the rank in the other list. The value after the name is the centrality score.

1	conrad, patricia, 0.123202
2	folkers, greg, 0.094716
3	billet, courtney, 0.075710
4	routh, jennifer, 0.064661
5	stover, kathy, 0.061491
6	marston, hilary, 0.056775
7	haskins, melinda, 0.043479
8	tabak, lawrence, 0.043263
9	fauci, anthony, 0.037303
10	mascola, john, 0.034584

Table 6: Top 10 nodes in terms of CC-based tensor H-eigenvector centrality. This is the only list in this document where Haskins and Mascola are top centrality nodes.

for all indices u and some scalar $\lambda > 0$. Since the first index of T corresponds to CC, the centrality scores are a measure of how central each node is with respect to participation in the CC role (x would be the same if we permuted the second and third indices, so only the first index determines the interpretation of the centrality).

Table 6 reports the top-10 nodes in terms of this centrality measure. Fauci is ranked ninth even though the entire dataset is constructed from his emails. However, Fauci is in the CC position relatively less often (Fauci was ranked first if the first index of the tensor corresponded to the sender or recipient roles). Conrad is ranked first, which agrees with her central role in many graphs constructed from this dataset. Folkers, Fauci's Chief of Staff, is ranked second. This tensor-based approach provides a way to highlight other notions of centrality not captured by our previous analyses. We find for example that this is the only centrality measure we consider where Melinda Haskins and John Mascola are both top ten centrality nodes.

Conclusions and Discussion

We have released an easy-to-use `json` file of the 3234-page PDF of Fauci's emails sent during early months of the COVID-19 pandemic in the Unites States. It is very likely that additional relevant correspondence took place over the phone and text messages that are not included in the data. Please also remember that this not *all* of Fauci's email from the relevant timeframe.

The processing of this data was automated. While we attempt to describe the major scenarios and edge-cases above and discuss how we handled them, please be aware that the dataset may contains some errors. In terms of sociological findings for which they may be appropriate, these data should be used with care to understand nuances regarding the exact data collection and ingestion. Note also that the text fields of our released data have many OCR errors. This

renders text analysis problematic and we leave text analysis to future studies.

Overall, we found this data extremely interesting for its seemingly unique ability to show differences among closely related methods. We have highlighted several of those features here and many more in an extended manuscript (Benson, Veldt, and Gleich 2021). The data is also small and easy-to-process, even with combinatorial optimization tools that are infeasible on larger data. We hope it becomes a useful resource to others as well!

References

Benson, A.; Veldt, N.; and Gleich, D. F. 2021. fauci-email: a json digest of Anthony Fauci's released emails. *arXiv*, cs.SI: 2108.01239. Code and data available from https://github.com/nveldt/fauci-email.

Benson, A. R. 2019. Three hypergraph eigenvector centralities. *SIAM Journal on Mathematics of Data Science*, 1(2): 293–312.

Benson, A. R.; Abebe, R.; Schaub, M. T.; Jadbabaie, A.; and Kleinberg, J. 2018. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*.

Bettendorf, N.; and Leopold, J. 2021. Anthony Fauci's Emails Reveal The Pressure That Fell On One Man. https://www.buzzfeednews.com/article/nataliebettendorf/fauci-emails-covid-response. Accessed: 2022-04-06.

Bhadra, S.; and Ferreira, A. 2003. Complexity of Connected Components in Evolving Graphs and the Computation of Multicast Trees in Dynamic Networks. In Pierre, S.; Barbeau, M.; and Kranakis, E., eds., *Ad-Hoc, Mobile, and Wireless Networks*, volume 2865 of *Lecture Notes in Computer Science*, 259–270. Springer Berlin / Heidelberg. ISBN 978-3-540-20260-8.

Cohen, W. W. 2004. Enron email dataset. http://www.cs.cmu.edu/~enron. Accessed: 2022-04-06.

Csardi, G.; and Nepusz, T. 2006. The igraph software package for complex network research. *InterJournal*, Complex Systems: 1695.

Freeman, L. C. 1977. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1): pp. 35–41.

Grindrod, P.; Parsons, M. C.; Higham, D. J.; and Estrada, E. 2011. Communicability across evolving networks. *Physical Review E*, 83(4).

Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2007. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1).

Liu, M.; Veldt, N.; Song, H.; Li, P.; and Gleich, D. F. 2021. Strongly Local Hypergraph Diffusions for Clustering and Semi-Supervised Learning. In *Proceedings of the Web Conference 2021*, WWW '21, 2092–2103. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mucha, P. J.; Richardson, T.; Macon, K.; Porter, M. A.; and Onnela, J.-P. 2010. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science*, 328(5980): 876–878.

Newman, M. E. J.; and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69(2): 026113.

Nicosia, V.; Tang, J.; Musolesi, M.; Russo, G.; Mascolo, C.; and Latora, V. 2012. Components in time-varying graphs. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(2): 023101.

Paranjape, A.; Benson, A. R.; and Leskovec, J. 2017. Motifs in Temporal Networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17. ACM.

Yin, H.; Benson, A. R.; Leskovec, J.; and Gleich, D. F. 2017. Local Higher-Order Graph Clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press.

Zachary, W. W. 1977. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4): 452–473.