AFRL-RI-RS-TR-2017-234

XDATA

SOTERA DEFENSE SOLUTIONS

DECEMBER 2017

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2017-234 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /	**/ S /**
NANCY ROBERTS	MICHAEL J. WESSING
Work Unit Manager	Deputy Chief, Information Intelligence
	Systems and Analysis Division
	Information Directorate

REPORT DOCUMENTATION PAGE	***Form Approved*** **OMB No. 0704-0188**

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* DECEMBER 2017	**2. REPORT TYPE** FINAL TECHNICAL REPORT	**3. DATES COVERED** *(From - To)* SEP 2012 – MAY 2017

4. TITLE AND SUBTITLE XDATA	**5a. CONTRACT NUMBER** FA8750-12-C-0301
	5b. GRANT NUMBER N/A
	5c. PROGRAM ELEMENT NUMBER 62702E
6. AUTHOR(S) Christopher Chabalko	**5d. PROJECT NUMBER** XDAT
	5e. TASK NUMBER A0
	5f. WORK UNIT NUMBER 10

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Sotera Defense Solutions 2121 Cooperative Dr., #400 Herndon, VA 20171	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RIEA 525 Brooks Road Rome NY 13441-4505	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFRL/RI
	11. SPONSOR/MONITOR'S REPORT NUMBER AFRL-RI-RS-TR-2017-234

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. PA# 88ABW-2017-6028
Date Cleared: 30 NOV 2017

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Data analysis tools which operate on varied data sources including time series, social media data, and e-mail data are described. These tools were the primary products and address the goal of furthering the field of data analysis. Through the course of this effort, an emphasis was placed on analysis over multiple scales from individual data elements to aggregates of millions of elements and relationships to other datasets. One particularly successful aspect of this effort was the development of an e-mail analysis program referred to as Newman. Newman provides a top down view of e-mail data sets as well as fast access to individual emails, attachments, and other aspects of e-mail correspondence.

15. SUBJECT TERMS

Big Data, XDATA, ETL, Newman Distributed Graph Analytics

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** NANCY ROBERTS
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	19	**19b. TELEPHONE NUMBER (*Include area code)***

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Major contributions to this research effort were made by Justin Gawrilow, Christopher Chabalko, Eli Fajardo, Joe Gartner, Jakob Lee, Elliot Ickovic, Michael Frame, Justin Lueders, John Reeme, and Kathleen Lossau, with additional guidance by James Lobb.

In addition, this effort was enhanced by the valuable guidance and discussions provided by DARPA’s systems engineering and technical advisors as well as the Program Manager.

The Sotera contract under the DARPA XDATA Program was sponsored and administered by the Air Force Research Laboratory (AFRL).

1.0 SUMMARY

Through the course of the XDATA program, Sotera has successfully worked with many types of data on local and distributed systems. In addition, portable cloud based infrastructure was created which was capable of operating on data sets which were intractable for a single machine.

Although the general format of data is too diverse for a single set of algorithms to successfully interact with, relationships between data elements have been identified as a critical value added attribute of data. The development of network graph algorithms to analyze and visualize these relationships provides analysts with a key tool in the dissection and understanding of varied data sets. In addition, the values and relationships of electronic mail (e-mail) data, which were previously cumbersome to identify, are revealed in a simple, fast, and scalable package, referred to as Newman, developed in the course of this effort. Newman makes the contents of e-mail and attachments easily searchable and also reveals the communications network from which the e-mail was generated. This data heavy application merges computational capability with the business use case of understanding human behaviors.

2.0 INTRODUCTION

Data continues to be generated and digitally archived at increasing rates, resulting in vast databases available for search and analysis. Access to these databases has generated new insights through data-driven methods in the commerce, science, and computing sectors. The so-call "big data" problem has now become a challenge for military operations, both at the strategic and tactical levels. The data being brought to bear on operations are growing rapidly in volume and complexity, and are most often imperfect, incomplete, heterogeneous, and consumed by diverse end-users from analysts to field soldiers.

The overall approach for the XDATA program is to develop open source software toolkits that enable flexible software development supporting users processing large volumes of data in timelines commensurate with mission workflows of targeted defense applications.

Feature correlation across multiple data sources, at ultra-large scales, is a tractable problem when performed within a framework on scalable architectures. Performing such analysis across large sets of heterogeneous data (such as a Global Graph) can bring robust, adaptable, and modular solutions to the newest and most aggressive types of modern intelligence analysis problems. This capability will provide the ability to identify patterns and anomalies within patterns based on all dimensions of the data rather than relying on subjective concepts of which set of dimensions are relevant. Our innovative approach employs multi-dimensional modeling implemented in cloud storage.

Although digital data is stored and transmitted in a variety of formats, many of these formats can be interpreted by open source tools or read directly to text, image, JavaScript object notation (JSON), or some other interchangeable format. Certain data elements, such as e-mail, also exist in a relational sense. In the case of e-mail, the message itself contains the relational information in the "to", carbon copy (cc), blind carbon copy (bcc), from, and attachment information. In the case of multiple similar pictures of the same event, the relationships are determined by considering many images and making comparisons between them. In each case, algorithms must be developed to facilitate interacting with and determining relationships between data elements.

The focus of this effort was to develop, test, and deliver automated and semi-automated computational analytical techniques, tools, and software libraries to discover statistically significant anomalies and patterns within very large data sets. Included in this are distributed graph analytic processes, cloud-based approximation mechanisms, and rapid development environment toolsets. Deliverables included monthly status reports, quarterly status reports, attendance at quarterly primary investigator (PI) meetings (hackathons), and developed software products.

3.0 METHODS, ASSUMPTIONS AND PROCEDURES

Computational techniques and software tools for analyzing large volumes of data, both semi-structured (e.g., tabular, relational, categorical, meta-data) and unstructured (e.g., text documents, message traffic) were developed. Central challenges included developing scalable algorithms for processing imperfect data in distributed data stores. In addition, e-mail data was identified as a rich data source, with multiple types of content (text, images, exchangeable image format (EXIF) data, header data) as well as relational traits. E-mail is also valuable as a log of human and organizational activities and relationships.

3.1 Infrastructure Development

Infrastructure was developed which integrated dimensional analytics, graph analytics, and portable cloud based development environments. A mechanism to move data and execution between cloud instances was researched and benchmarked. Finally, a visualization component, designed as middleware, was developed.

3.2 Data Representation

Several data representations were researched throughout the course of the XDATA program. This investigation positively impacted many of the tools described in the Transitions and Demonstrations section below. The goals of the investigations included:

- Evaluation of schema alignment techniques such as combined matching (COMA)++
- Research, employ and leverage semi-automated and automated data cleansing techniques and report on how these capabilities can be integrated and used within the XDATA architecture
- Provide data normalization and representation documentation for the provided extract, transform, and load (ETL) tools and algorithms
- Research and provide evaluation reports on data optimization techniques

4.0 RESULTS AND DISCUSSION

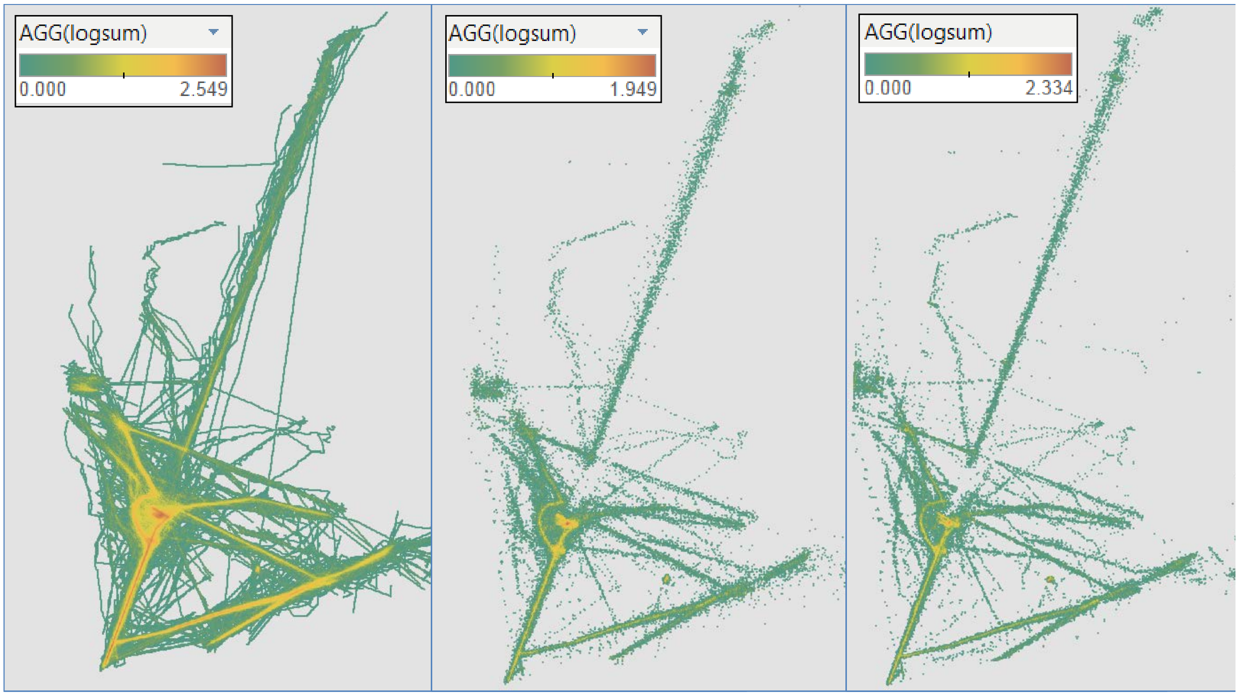
The contractor participated in all of the XDATA summer camps and activities as well as the quarterly PI meetings referred to as hackathons. Many products were developed to the research level during the course of this contract. Several of these are described below.

4.1 Aggregate micro paths

Links to the github repository for the aggregate micro path effort are provided as: http://sotera.github.io/aggregate-micro-paths/, and https://github.com/Sotera/aggregate-micro-paths.

The geo-location in Twitter based social media tweets can provide an overview of areas which are commonly navigated by humans. For example, in a city, the geo-location of tweets may provide an overview of locations where pedestrians are commonly present, such as sidewalk and pedestrian paths. (A prototype dataset used and is no longer available.) Through the course of the XDATA program, the geo-information was considered in aggregate to identify micro-paths, not otherwise identified in the individual observations. Rather than simply identifying the presence of a sidewalk, the micro-paths identify commonly traveled routes. The routes enable the ability to predict a travel pattern given only a single or few tweets. Since the geo-information is contained completely in the twitter set, route projections can be performed in the absence of a base map, and in the absence of any prior knowledge about popular routes, attractions, or other pedestrian objectives.

The difference between raw detections, raw tracks, and micropaths is shown in Figure 1. Commonly traveled routes are shown in red, while less commonly traveled routes are show in progressively greener colors. On well traveled portions of the route, the raw detections are dense and provide a distinct view of the boarder of the path. On less well traveled portions of the terrain, the tweets are sparse, however taken in aggregate they provide a view of a less popular route. Automated construction of micropaths requires cleaning the data points as well as considering multiple scales ranging from single observations to large scale complete paths.



Micro-paths (10M)
(No distance/time filter)

Raw Tracks (10M)

Raw Detections (10M)

Figure 1. Micro-paths, raw tracks, and raw detections from geo-located tweets

4.2 Correlation Approximation

A link to the github repository for the Correlation Approximation effort is provided as: https://github.com/Sotera/correlation-approximation.

Provided with experimental observations, the identification of correlations can often led to insights into the governing mechanisms of an observed behavior. Identification of correlations are especially valuable when multiple experiments are coupled by a common factor such as human behavior. While correlation does not necessarily imply causation, many valuable insights have been gained by simply identifying correlations. One anecdotal example is the correlation between the purchase beer and diapers. While this anecdotal correlation may not have been exploited to increase retail sales, other retail correlations certainly could have. This powerful tool can extend to many aspects of marketing, intelligence gathering, and socio-economic effects. Google Correlate was developed specifically to identify and leverage useful but computationally expensive correlations. A similar functionality was achieved in this effort. The correlation procedure is computationally expensive as naive implementations scale with order (O) (N^2) and can be reduced to O(N*log(N)) through application in Fourier space. The correlation approximation engine, results of which are shown in Figure 2, applies an approximation to ensure low computational complexity. Furthermore, the algorithm is implemented in Spark to allow for implicit scalability.

Correlations are typically computed between only two variables. The correlation approximation engine computes the correlation between all variables in a given data set, excluding auto-correlations. For example, computing the cross correlation between 100 time series results in 4950 individual correlation computations. Each correlation is computationally expensive on its own, however, each computation is also independent from the others. This independence was leveraged to implement the code in the Spark distributed computing environment.

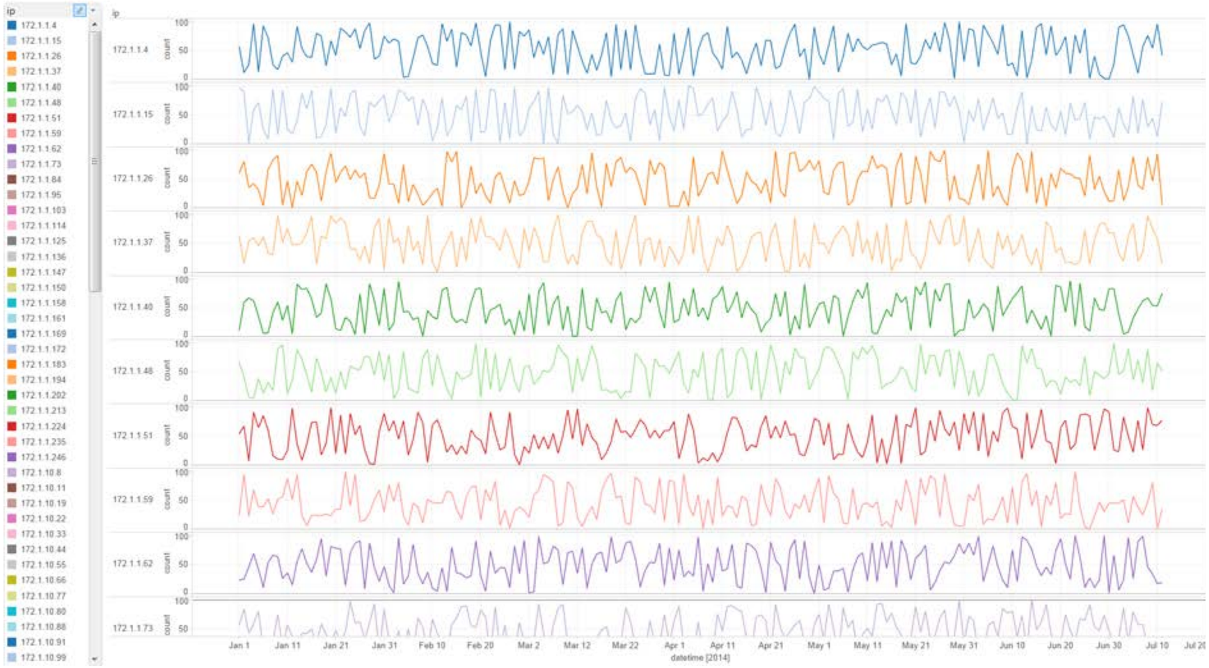


Figure 2. Example from the correlation approximation engine

4.3 Distributed Graph Analytics

Links to the github repository for the Distributed Graph Analytics effort are provided as: http://sotera.github.io/distributed-graph-analytics/, and https://github.com/Sotera/distributed-graph-analytics.

Distributed Graph Analytics (DGA) is a compendium of graph analytics written for Bulk-Synchronous-Parallel (BSP) processing frameworks such as Giraph and GraphX. The DGA toolset allows parallel computation of network graph algorithms including: High Betweenness Set Extraction, Weakly Connected Components, Page Rank, Leaf Compression, and Louvain Modularity. Not only are computations performed in parallel, but the Hadoop file systems is utilized for operation on large graphs which may otherwise exceed the RAM resources on a typical machine.

Louvain modularity, implemented in DGA, can be used to identify clusters in a network graph, as shown in Figure 3. This is a critical step for visualizing and identifying macroscopic trends in certain types of network graphs. DGA was benchmarked against the following data sets:

- Bitcoin transaction graph 2,132,321 vertices 25 days to calculate exact betweenness on a single core. (A prototype dataset used and is no longer available.)

- Brightkite social network. 58,228 vertices, 18 minutes to calculate exact betweenness on a single core. (Leskovec & Krel, Brightkite, 2014)
- Enron e-mail data set. 35,818 vertices, 156 seconds to calculate exact betweenness on a single core. (Enron dataset, 2013)
- Gowalla social network. 196,591 vertices (not calculated on single core) (Leskovec & Krevl, Gowalla, 2014)
- Google Plus social network. 102,118 vertices, 14.5 hours to calculate exact betweenness on a single core. (Leskovec & Krevl, Google+, 2014)

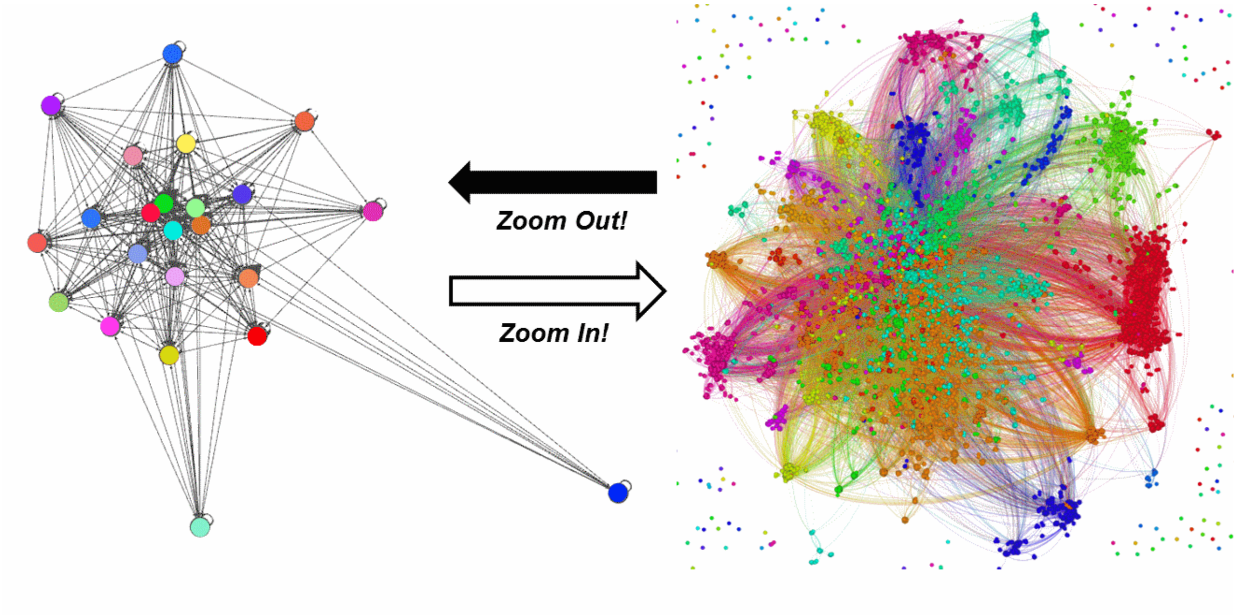


Figure 3. Louvain modularity is used to identify significant features of a network graph

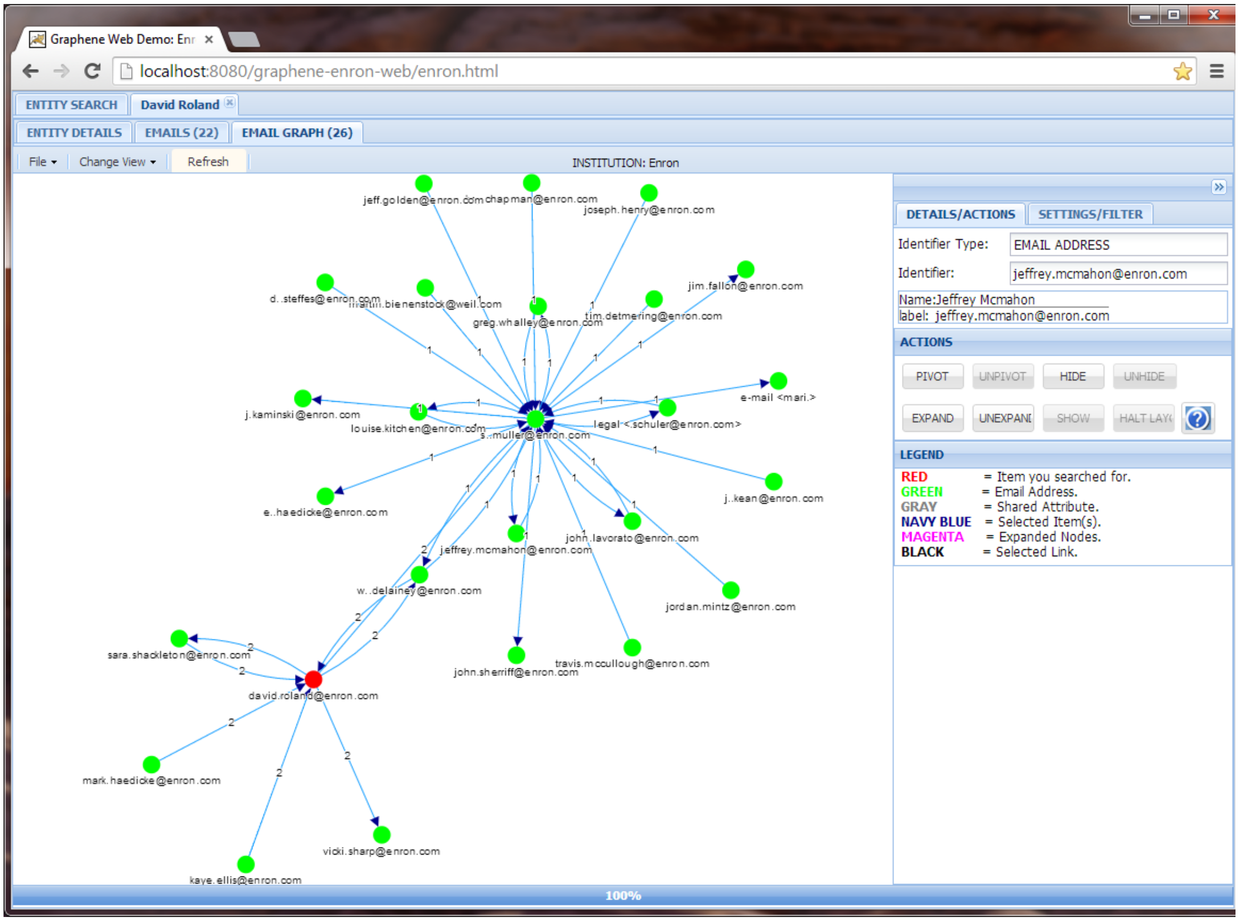
In each case, the DGA implementation significantly accelerated the network graph analysis. The DGA tool also exposes parameters which can be varied to provide approximate results rapidly. The highest accuracy results for each data set, as computed with an 8-node cluster using 8 giraph workers using 8 threads per worker were:

- Bitcoin transaction graph ~3 hours at 87% accuracy -> 200x speedup
- Brightkite social network ~1.5 minutes at 81% accuracy -> 12x speedup
- Enron e-mail data set ~ 1 minute at 65% accuracy -> 2.6x speedup
- Gowalla social network ~ 6 minutes at 75% accuracy -> not available (N/A)
- Google Plus social network ~15 minutes @ 81.2% accuracy -> 58.3x speedup

4.4 Graphene

A link to the github repository for Graphene is provided as: https://github.com/Sotera/graphene

Graphene is a high performance Java based web framework used to build a searching and graphing application on top of existing data, as shown in Figure 4. It is datastore agnostic, but has built in support for Elastic Search, structured query language (SQL) Databases, and Titan. The relational view of data elements is beneficial for many investigations including financial transactions.



Identifier Type: EMAIL ADDRESS

Identifier: jeffrey.mcmahon@enron.com

Name:Jeffrey Mcmahon
label: jeffrey.mcmahon@enron.com

RED = Item you searched for.
GREEN = Email Address.
GRAY = Shared Attribute.
NAVY BLUE = Selected Item(s).
MAGENTA = Expanded Nodes.
BLACK = Selected Link.

Figure 4. Graphene showing a network graph view of a dataset

4.5 Newman

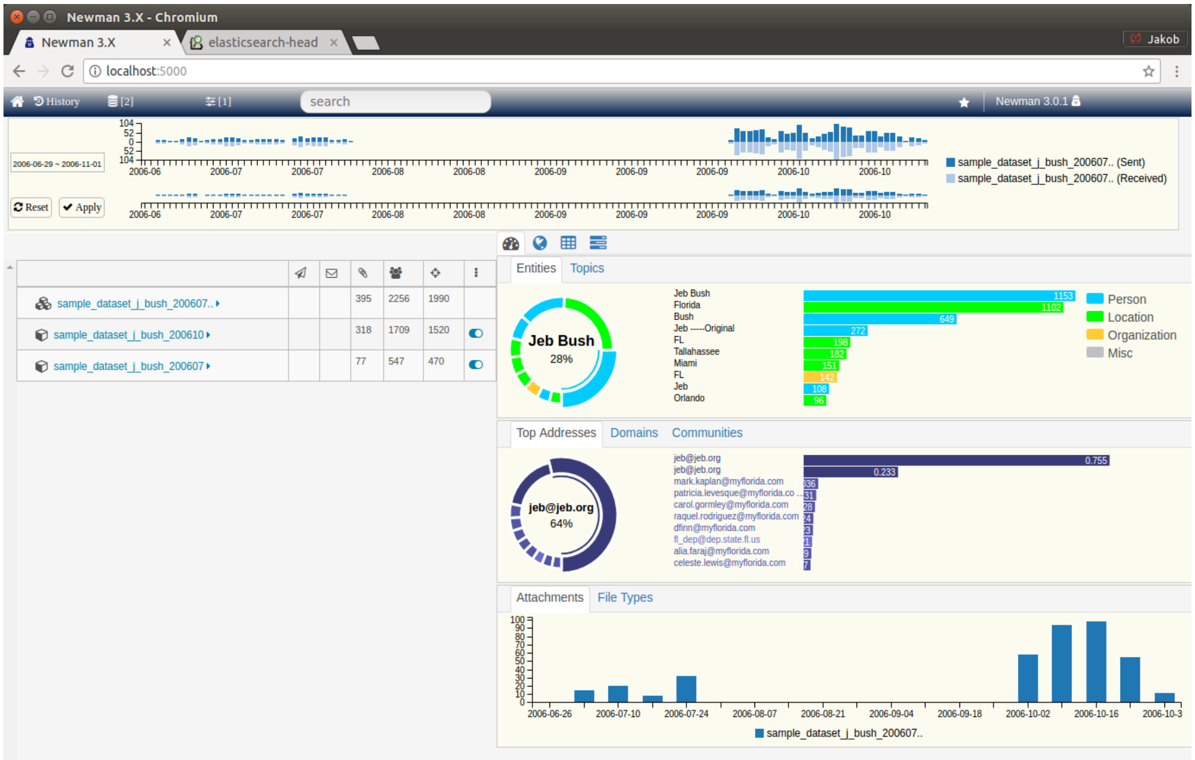
A link to the github repository for Newman is provided as: https://github.com/Sotera/newman.

Newman is an e-mail analysis and exploitation program. Newman provides an overview of e-mails in a data set as well as detailed views of e-mail contents, geographic origin of images and internet protocol (IP) addresses, searchable attachments and much more information.

The e-mail exploitation process begins with the Newman ETL process. After one button ingest is triggered, e-mail data sets are ingested. In this process, e-mail text, attachment text, including text contained in word documents, .pdf's, excel sheets, and many other formats is indexed in an elastic search database. Language translation and optical character recognition is performed on ingest as well.

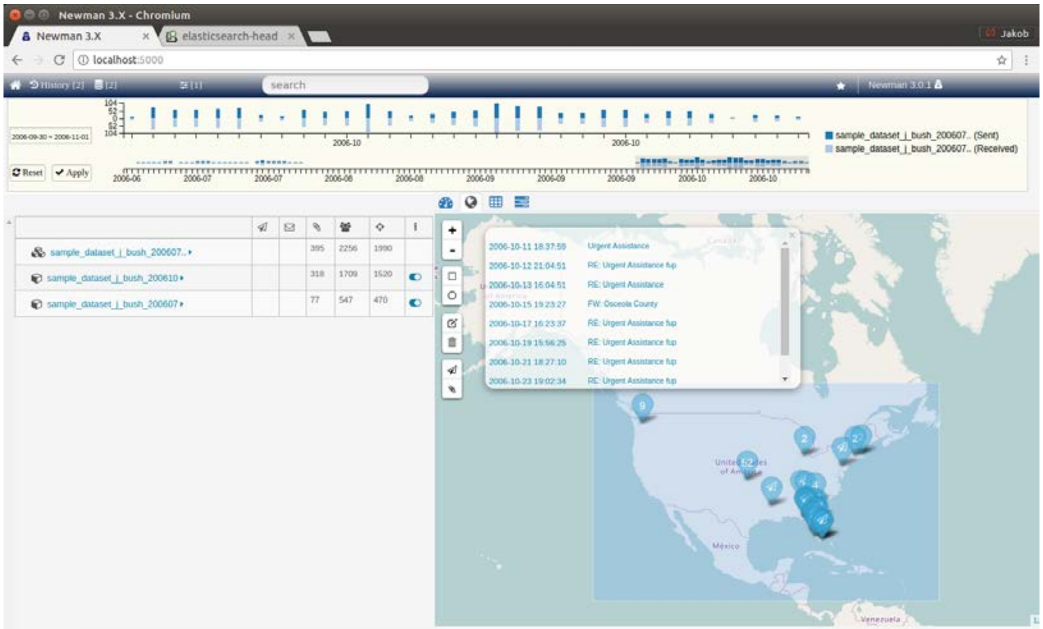
Next the analyst is presented with an overview of the e-mail data set (American Bridge 21st Century, n.d.). The overview includes histograms of e-mails sent and received over time, topic extraction, summarization, and a characterization of the distribution of attachment types as seen in Figure 5. E-mail accounts can be optionally enabled/disabled so analysts can focus on a single account or group of accounts, rather than all accounts.

The search box near the top allows access to Elasticsearch's search capabilities including: fuzzy search, exact matching, Boolean search combination, regular expressions, and more. Many elements of the dashboard are active in the sense that they allow filtering of the data set or a more detailed view of some particular aspect of the data set. Both image EXIF data and e-mail IP addresses can be used to locate elements on a map as shown in Figure 6.



sample_dataset_j_bush_200607.. ▸			395	2256	1990	
sample_dataset_j_bush_200610 ▸			318	1709	1520	●
sample_dataset_j_bush_200607 ▸			77	547	470	●

Figure 5. Newman overview screen

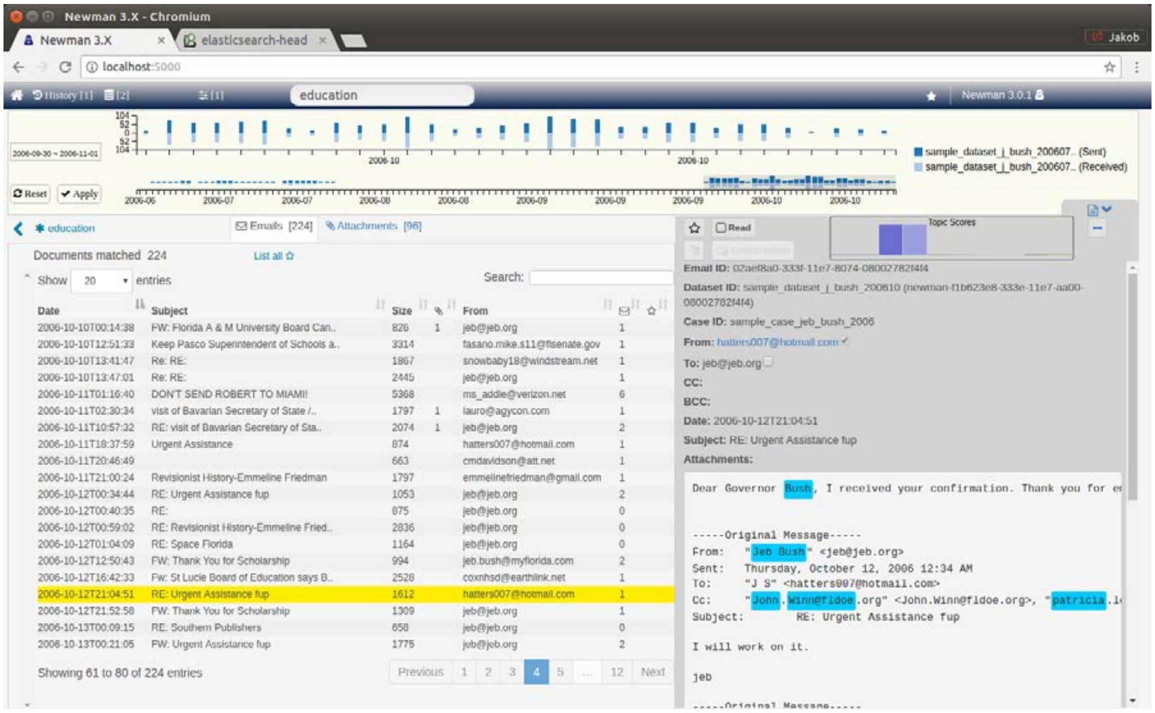


sample_dataset_j_bush_200607.. ▸			395	2256	1990	
sample_dataset_j_bush_200610 ▸			318	1709	1520	●
sample_dataset_j_bush_200607 ▸			77	547	470	●

2006-10-11 18:37:59	Urgent Assistance
2006-10-12 21:04:51	RE: Urgent Assistance fup
2006-10-13 16:04:51	RE: Urgent Assistance
2006-10-15 19:23:27	FW: Osceola County
2006-10-17 16:23:37	RE: Urgent Assistance fup
2006-10-19 15:56:25	RE: Urgent Assistance fup
2006-10-21 18:27:10	RE: Urgent Assistance fup
2006-10-23 19:02:34	RE: Urgent Assistance fup

Figure 7. Map view showing geo-location of images and e-mail IP addresses

E-mail contents can be viewed in the detail view as shown in Figure 7. Search terms and extracted entities are highlighted in this view as well. Rather than e-mail text, a view of all attachments is available by clicking on the attachments tab near the top of the results window as shown in Figure 8.



Documents matched 224 List all ☆

Show 20 entries Search:

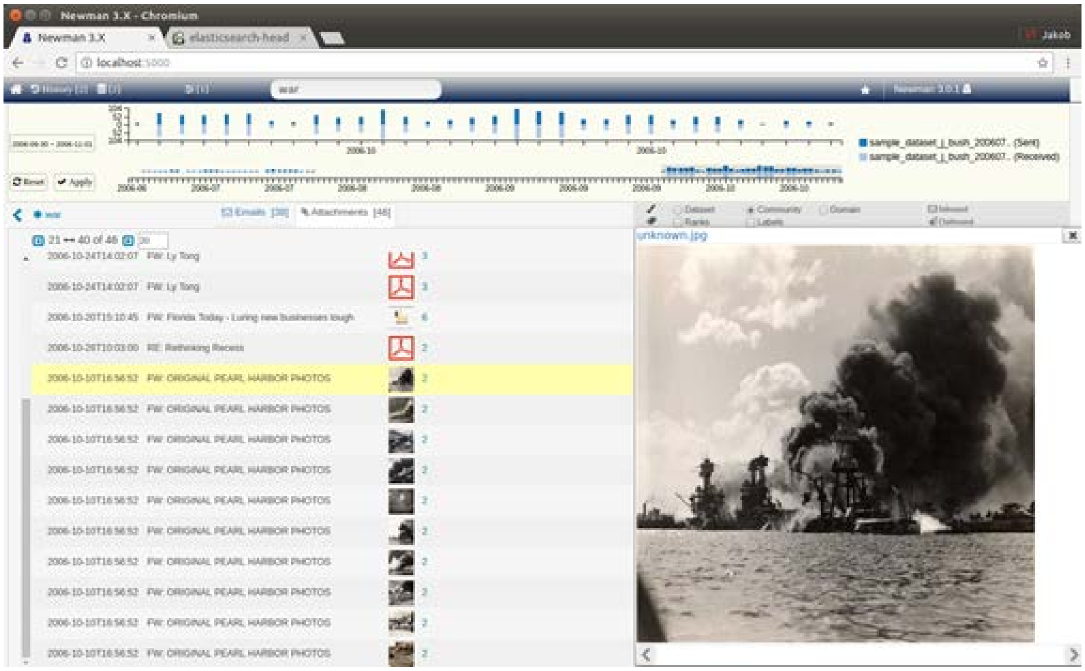
Date	Subject	Size	✎	From	✉	☆
2006-10-10T00:14:38	FW: Florida A & M University Board Can..	826	1	jeb@jeb.org	1	
2006-10-10T12:51:33	Keep Pasco Superintendent of Schools a..	3314		fasano.mike.s11@flsenate.gov	1	
2006-10-10T13:41:47	Re: RE:	1867		snowbaby18@windstream.net	1	
2006-10-10T13:47:01	Re: RE:	2445		jeb@jeb.org	1	
2006-10-11T01:16:40	DON'T SEND ROBERT TO MIAMI!	5368		ms_addie@verizon.net	6	
2006-10-11T02:30:34	visit of Bavarian Secretary of State /..	1797	1	lauro@agycon.com	1	
2006-10-11T10:57:32	RE: visit of Bavarian Secretary of Sta..	2074	1	jeb@jeb.org	2	
2006-10-11T18:37:59	Urgent Assistance	874		hatters007@hotmail.com	1	
2006-10-11T20:46:49		663		cmdavidson@att.net	1	
2006-10-11T21:00:24	Revisionist History-Emmeline Friedman	1797		emmelinefriedman@gmail.com	1	
2006-10-12T00:34:44	RE: Urgent Assistance fup	1053		jeb@jeb.org	2	
2006-10-12T00:40:35	RE:	875		jeb@jeb.org	0	
2006-10-12T00:59:02	RE: Revisionist History-Emmeline Fried..	2836		jeb@jeb.org	0	
2006-10-12T01:04:09	RE: Space Florida	1164		jeb@jeb.org	0	
2006-10-12T12:50:43	FW: Thank You for Scholarship	994		jeb.bush@myflorida.com	2	
2006-10-12T16:42:33	Fw: St Lucie Board of Education says B..	2528		coxnhsd@earthlink.net	1	
2006-10-12T21:04:51	RE: Urgent Assistance fup	1612		hatters007@hotmail.com	1	
2006-10-12T21:52:58	FW: Thank You for Scholarship	1309		jeb@jeb.org	1	
2006-10-13T00:09:15	RE: Southern Publishers	658		jeb@jeb.org	0	
2006-10-13T00:21:05	FW: Urgent Assistance fup	1775		jeb@jeb.org	2	

Showing 61 to 80 of 224 entries   Previous 1 2 3 4 5 … 12 Next

Email ID: 02aef8a0-333f-11e7-8074-08002782f4f4
Dataset ID: sample_dataset_j_bush_200610 (newman-f1b623e8-333e-11e7-aa00-08002782f4f4)
Case ID: sample_case_jeb_bush_2006
From: hatters007@hotmail.com
To: jeb@jeb.org
CC:
BCC:
Date: 2006-10-12T21:04:51
Subject: RE: Urgent Assistance fup
Attachments:

```
Dear Governor Bush, I received your confirmation. Thank you for e


-----Original Message-----
From:    "Jeb Bush" <jeb@jeb.org>
Sent:    Thursday, October 12, 2006 12:34 AM
To:      "J S" <hatters007@hotmail.com>
Cc:      "John.Winn@fldoe.org" <John.Winn@fldoe.org>, "patricia.l
Subject:        RE: Urgent Assistance fup

I will work on it.

jeb

-----Original Message-----
```

Figure 6. Newman: detailed view

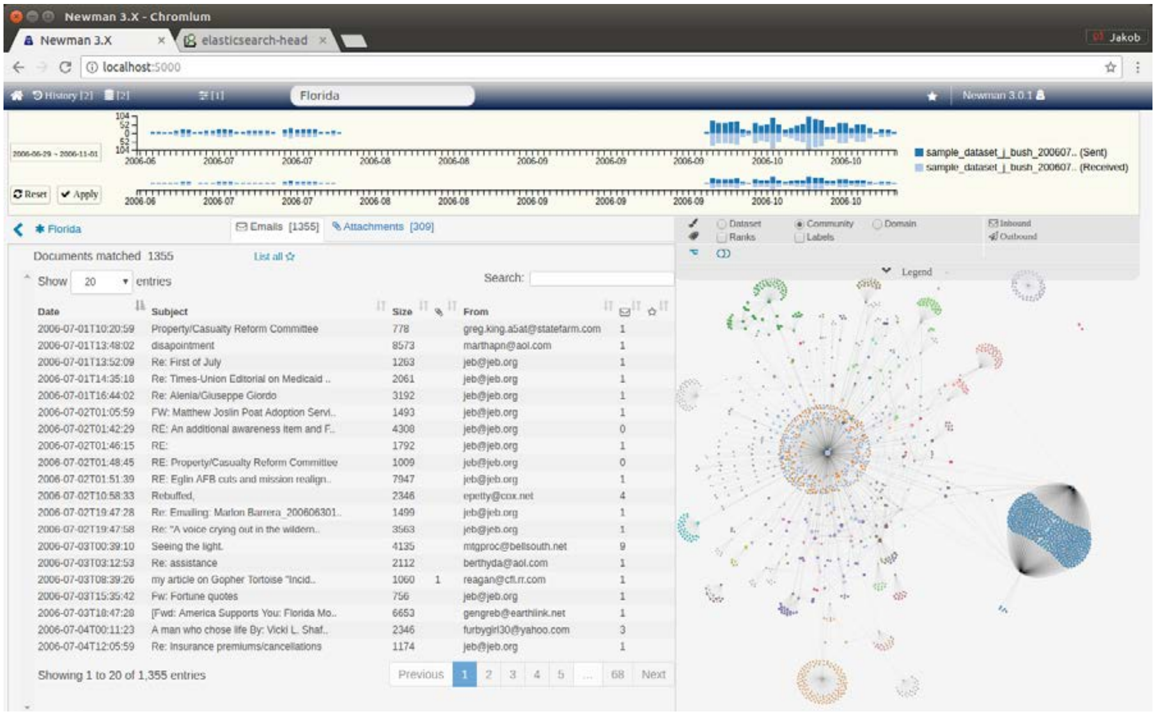


21 ↔ 40 of 46

2006-10-24T14:02:07	FW: Ly Tong	3
2006-10-24T14:02:07	FW: Ly Tong	3
2006-10-20T15:10:45	FW: Florida Today - Luring new businesses tough	6
2006-10-20T10:03:00	RE: Rethinking Recess	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2
2006-10-10T16:56:52	FW: ORIGINAL PEARL HARBOR PHOTOS	2

Figure 8. Newman attachments view

Finally, a network view of e-mail correspondence is automatically generated. The network nodes can be highlighted based on the computed community, the dataset membership, or e-mail domain, as shown in Figure 9.



Documents matched 1355 List all ☆

Show 20 entries Search:

Date	Subject	Size	✎	From	✉	☆
2006-07-01T10:20:59	Property/Casualty Reform Committee	778		greg.king.a5at@statefarm.com	1	
2006-07-01T13:48:02	disapointment	8573		marthapn@aol.com	1	
2006-07-01T13:52:09	Re: First of July	1263		jeb@jeb.org	1	
2006-07-01T14:35:18	Re: Times-Union Editorial on Medicaid ..	2061		jeb@jeb.org	1	
2006-07-01T16:44:02	Re: Alenia/Giuseppe Giordo	3192		jeb@jeb.org	1	
2006-07-02T01:05:59	FW: Matthew Joslin Poat Adoption Servi..	1493		jeb@jeb.org	1	
2006-07-02T01:42:29	RE: An additional awareness item and F..	4308		jeb@jeb.org	0	
2006-07-02T01:46:15	RE:	1792		jeb@jeb.org	1	
2006-07-02T01:48:45	RE: Property/Casualty Reform Committee	1009		jeb@jeb.org	0	
2006-07-02T01:51:39	RE: Eglin AFB cuts and mission realign..	7947		jeb@jeb.org	1	
2006-07-02T10:58:33	Rebuffed,	2346		epetty@cox.net	4	
2006-07-02T19:47:28	Re: Emailing: Marlon Barrera_200606301..	1499		jeb@jeb.org	1	
2006-07-02T19:47:58	Re: "A voice crying out in the wildern..	3563		jeb@jeb.org	1	
2006-07-03T00:39:10	Seeing the light.	4135		mtgproc@bellsouth.net	9	
2006-07-03T03:12:53	Re: assistance	2112		berthyda@aol.com	1	
2006-07-03T08:39:26	my article on Gopher Tortoise "Incid..	1060	1	reagan@cfl.rr.com	1	
2006-07-03T15:35:42	Fw: Fortune quotes	756		jeb@jeb.org	1	
2006-07-03T18:47:28	[Fwd: America Supports You: Florida Mo..	6653		gengreb@earthlink.net	1	
2006-07-04T00:11:23	A man who chose life By: Vicki L. Shaf..	2346		furbygirl30@yahoo.com	3	
2006-07-04T12:05:59	Re: Insurance premiums/cancellations	1174		jeb@jeb.org	1	

Showing 1 to 20 of 1,355 entries Previous 1 2 3 4 5 … 68 Next

Figure 9. Newman, communications network graph view

The network nodes represent senders or receivers of e-mails. Connections indicate that a communication has taken place between the entities. Many of the network graph computational algorithms used to compute quantities related to the network graph were also developed as a part of the XDATA program.

5.0 CONCLUSIONS

The XDATA program served as a foundry to create tools which can interact with and provide analysis of many types of data. While this goal may seem abstract, our team made substantial contributions through the development of three primary types of systems. First, many of the tools operate at multiple scales simultaneously. In the case of micro-paths, a bottom up approach which considered individual tweets first then aggregated to address larger scale features was applied. In the case of e-mail analysis, a top down approach which considered an overview of the corpus with the goal of moving to smaller scale high value elements (e-mail text and attachments) was executed. Finally, the computational demands of operating over large and small scales simultaneously were addressed through the development of distributed systems. The virtualized cloud infrastructure developed in this effort filled a similar role to that of commercial cloud computing providers, although several years earlier. In conclusion, substantial benefits can be gain from considering local data values as well as data in relation to other, possible very different, data sets.

6.0 RECOMMENDATIONS

We have analyzed a variety of data sources and processing techniques through the course of the XDATA program. Many of our most useful efforts were enabled by the application of open source software on diverse data sets. This was most clear during the PI meetings (hackathons). In these cases, diverse datasets were provided, with analysis goals, but little specific instructions on how to achieve those goals. The freedom to analyze the data and bring in outside data sources resulted in value and insight which was not anticipated at the start of the program. Having access to open source tools as well as diverse data sources appeared to greatly facilitate the work performed.

While algorithms can be packaged and refined, the ETL process is different for every data set. This process consumes the majority of the time spent in processing a new data. An automated ETL and dimensionality assessment of a dataset would accelerate the application of data science and allow a more diverse audience to enjoy the fruits of programs such as XDATA.

BIBLIOGRAPHY

Retrieved from American Bridge 21st Century: https://americanbridgepac.org/jeb-bushs-gubernatorial-email-archive/

Enron dataset. (2013, May 15). Retrieved from Nuix website: http://info.nuix.com/Enron.html

Leskovec, J. (n.d.). *Brightkite*. Retrieved from SNAP Datasets: Large Network Dataset Collection: https://snap.stanford.edu/data/loc-brightkite.html

Leskovec, J., & Krevl, A. (2014, Jun). *Google+*. Retrieved from SNAP Datasets: Stanford Large Network Dataset Collection: https://snap.stanford.edu/data/egonets-Gplus.html

Leskovec, J., & Krevl, A. (2014, Jun). *Gowalla*. Retrieved from SNAP Datasets: Stanford Large Network Dataset Collection: http://snap.stanford.edu/data/loc-gowalla.html

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

e-mail	**E**lectronic **mail**
JSON	**J**ava**S**cript **o**bject **n**otation
Cc	**C**arbon **c**opy
Bcc	**B**lind **c**arbon **c**opy
PI	**P**rincipal **i**nvestigator
EXIF	**Ex**changeable **i**mage **f**ormat
COMA	**Co**mbined **ma**tching
ETL	**E**xtract, **t**ransform, **l**oad
O	**O**rder
DGA	**D**istributed **G**raph **A**nalytics
N/A	**N**ot **a**vailable
SQL	**S**tructured **q**uery **l**anguage
IP	**I**nternet **p**rotocol
BSP	**B**ulk **S**ynchronous **P**arallel