Data Acquisition and Preparation for Social Network Analysis Based on Email

Lessons Learned

Fred Ma
Joint Studies Operational Research Team

Dave Allen
Experimentation Operational Research Team

Patrick Dooley
Canada Command Operational Research Team

DRDC CORA TM 2009-030
June 2009

Defence R&D Canada
Centre for Operational Research & Analysis

Joint Staff Operational Research Team

National Defence Défense nationale

Canada

Data Acquisition and Preparation for Social Network Analysis Based on Email

Lessons Learned

Fred Ma
Joint Studies Operational Research Team

Dave Allen
Experimentation Operational Research Team

Patrick Dooley
Canada Command Operational Research Team

Defence R&D Canada – CORA

Technical Memorandum
DRDC CORA TM 2009-030
June 2009

Principal Author

Original signed by Fred Ma

Fred Ma

Defence Scientist, DRDC CORA

Approved by

Original signed by Charles Morrisey

Charles Morrisey

Acting Section Head, Joint & Common OR

Approved for release by

Original signed by Dale Reding

Dale Reding

Chief Scientist, DRDC CORA

This work was sponsored by Canadian Forces Experimentation Centre, Command and Sense Team, under project Polar Guardian.

Defence R&D Canada – Centre for Operational Research and Analysis (CORA)

Abstract

In sharing information to improve situational awareness, other government departments and remotely situated outposts may vary in their reporting of information. A social network analysis was initiated within the Department of National Defence to show where informal communication may be significant to information sharing. The study was undertaken circa Q3 2006 by the Experimentation Operational Research Team at the Canadian Forces Experimentation Centre for the Command and Sense Team. Analytical results are not available, as the undertaking was not completed. This report describes the lessons learned in planning the data collection and preparation for the social network analysis.

The work was done under project Polar Guardian, the goal of which was to assess situational awareness in the arctic. The plan for the social network analysis included an initial email-based phase and a follow-up survey-based phase. This report focuses on the email phase; it is not a comparison of the two phases as separate approaches.

Due to the short time frame for conducting the trial on the social network analysis approach, in-house methods for data acquisition and analysis were explored. The main challenges in this approach arise from generating the communications data from email tracking logs in isolation from other information gathering and information providing parts of a corporate computer network.

Commercial tools were investigated, and warrant further examination. Their use requires a longer time frame for approval and installation on the Defence Wide Area Network.

Of the commercial and home-grown approaches, the most time is likely needed for solutions involving access to the servers, and deployment of applications on them.

Direct access to subject matter expertise in email administration is essential to arriving at a means for effective and timely data gathering and preparation. Such access is also essential for an interagency social network analysis, the issues of which are touched upon in this technical memorandum only at a high level.

Résumé

Dans le partage de l'information pour augmenter les connaissances de la situation, il peut y avoir des variations dans les rapports d'information produits par les autres ministères et les postes en région éloignée. Une analyse des réseaux sociaux a été entreprise au sein du Ministère de la Défense nationale pour démontrer comment les communications informelles peuvent avoir de l'importance dans le partage de cette information. L'équipe de recherche opérationnelle expérimentale du Centre d'expérimentation des Forces canadiennes a entrepris cette étude pour le compte de l'équipe Commandement et détection pendant le troisième trimestre de 2006. L'équipe n'a pas publié les résultats de l'analyse puisque l'étude a été suspendue. Le présent rapport décrit les leçons retenues dans la planification de la collecte de données et dans la préparation de l'analyse des réseaux sociaux.

Le travail a été exécuté dans le cadre du projet Polar Guardian, dont le but était d'évaluer les connaissances de la situation dans l'Arctique. Le plan de l'analyse des réseaux sociaux comprenait une première étape axée sur les courriels et une deuxième étape axée sur un sondage de suivi. Le présent rapport se penche sur l'étape axée sur les courriels. Il ne s'agit pas cependant d'une comparaison des deux étapes en tant qu'approches distinctes.

En raison du court délai pour mener les essais sur l'approche pour l'analyse des réseaux sociaux, nous avons étudié des méthodes internes pour acquérir et analyser les données. Les principaux défis de cette approche sont la génération des données de communication à partir des registres de suivi des courriels, de manière isolée des autres composantes de collecte et de partage d'information dans un réseau informatique d'entreprise.

Notre exploration préliminaire des outils commerciaux justifie un examen approfondi de ceux-ci. L'utilisation de ces outils nécessite cependant un long délai pour les approuver et les installer sur le Réseau étendu de la Défense.

De toutes les solutions commerciales ou internes étudiées, les méthodes qui prendront probablement le plus de temps sont celles qui nécessitent un accès aux serveurs et un déploiement d'applications sur les serveurs.

Un accès direct aux experts en la matière sur l'administration des courriels est un élément essentiel afin d'arriver à un moyen efficace de recueillir et de préparer les données en temps utile. L'accès à ce genre d'expertise est également essentiel à une analyse des réseaux sociaux entre les organisations, dont les défis sont couverts sommairement dans le présent document.

Executive summary

Data Acquisition and Preparation for Social Network Analysis Based on Email: Lessons Learned

Fred Ma; Dave Allen; Patrick Dooley; DRDC CORA TM 2009-030; Defence R&D Canada – CORA; June 2009.

Introduction: Arctic security is becoming an area of growing concern. In support of the Canadian Forces Experimentation Centre's (CFEC's) Command and Sense team[1], a social network analysis (SNA) was undertaken circa Q3 2006 by the CFEC's Experimentation Operational Research Team (EXORT) to provide visibility into the informal sharing of information between agencies and within the Department of National Defence (DND) that can affect situational awareness in the arctic. It was part of a larger project, Polar Guardian, whose purpose included (at different times) modelling 'As-Is' surveillance capabilities, identifying shortcomings, and modelling 'To-Be' capabilities to guide decisions on the way ahead. As a first step, an SNA was considered important because of the expected variability with which different organizations report information, especially in remote regions. A view of information communicated informally would give an idea of the accuracy of, and possibly augment, modelling of standard operating procedures for reporting.

Since interagency sharing of information is anything but trivial, an internal DND SNA was first undertaken on the Defence Wide Area Network. Due to changing priorities in CFEC's transformation, however, Polar Guardian was terminated in the data acquisition planning phase and the SNA was not performed. This technical memorandum captures lessons learned in the acquisition and preparation of data on corporate email, which comprises the first phase of the SNA; it is not a comparison of email- and survey-based SNA. The challenges to the former (not known at the outset) were driven by constraints in administration, policy, time, cost, and access to technical expertise. They differ from those for an SNA of communications in an experiment-specific common operating environment; the volume of data is larger, and issues arose from the fact that only tracking logs were accessible (pending legal approval, which was being pursued when Polar Guardian was put on hold).

Results: The amounts and forms of user identification data was highly variable, and results from efforts in characterizing the data are documented. Home-grown approaches to user identification are mapped out, along with their limitations, uncertainties, and challenges. The expedient approach of discarding irresolvable data would yield an SNA of unknown accuracy. Commercial tools for generating the data were vetted based on the constraints. Recommendations are given for future SNAs, including activities to start early in an SNA due to potentially long resolution times.

As an alternative to home-grown solutions, an investigation is suggested of the capabilities built into the mail server software, as is re-examination of commercial tools that provide summary statistics (possibly using directory services for user identification). These require a longer time

[1] CFEC has since reorganized into different teams.

frame for access/installation approvals. For commercial products, time is also needed to try the product out, and for purchasing administration.

Access to subject matter expertise in email administration and/or Microsoft Exchange Server® is essential, especially if it can include knowledge about email administration within DND. In addition to its requirement for solutions implemented on the mail servers, such experience narrows down the contingencies for which solutions need to be planned. It also informs the assessment of: (1) how the SNA is impacted by a solution's shortcomings in identification, (2) the challenges and risk/uncertainties (technical and administrative/policy), and (3) the resources and level of effort needed to accommodate or mitigate the challenges and uncertainties. For example, knowledge of the degree to which solutions encroach upon security-motivated restrictions can be taken into consideration in planning the implementation of those solutions; any measures that can be taken to minimize the encroachment improves the chances of approval for such solutions.

Significance: Interagency sharing of information involves communication between different organizations, with variability in training and culture (corporate and social). Compared to communication within DND, therefore, it is reasonable to expect a greater variation in adherence to formal procedures for prompt reporting of information, particularly in remote locations. An SNA can indicate where informal communications can be significant to situational awareness. The discovered challenges, potential approaches to their solution, and lessons captured here can inform the planning of future SNAs.

Future plans: For an email SNA within DND, the approaches scoped out in this document vary in detail. Some require further investigation, and a solution to user identification is yet to be implemented. For an interagency SNA, the detailed challenges are largely unknown, though anticipated challenges at a general level are presented. The analyst for such a study should work with subject matter experts to map out technical, policy, and cultural challenges and solutions. It is expected that a prior SNA within DND would help generate buy-in among other governmental departments (OGDs). Though there are trade-offs, surveys and interviews can be used exclusively if generating data from email is beyond the scope of the SNA[2].

[2] In terms of effort, resourcing, and time frame.

Sommaire

Data Acquisition and Preparation for Social Network Analysis Based on Email: Lessons Learned

Fred Ma; Dave Allen; Patrick Dooley; DRDC CORA TM 2009-030; R & D pour la défense Canada – CORA; Juin 2009.

Introduction : La sécurité dans l'Arctique est devenue un domaine de préoccupation croissant. En soutien à l'équipe Commandement et détection du Centre d'expérimentation des Forces canadiennes (CEFC)[3], l'équipe de recherche opérationnelle expérimentale (EXORT) du CEFC a entrepris une analyse des réseaux sociaux (ARS) pendant le troisième trimestre de 2006 afin de donner de la visibilité au partage d'information informel entre les organisations et au sein du ministère de la Défense nationale (MDN) pouvant avoir une importance sur les connaissances de la situation dans l'Arctique. Cette étude faisait partie d'un projet plus vaste, Polar Guardian, dont le but comprenait (à des moments différents) la modélisation des capacités de surveillance actuelle, la détermination des lacunes, et la modélisation des capacités voulues pour guider les décisions à venir. Comme première étape, on a jugé important d'exécuter une analyse des réseaux sociaux à cause des différences attendues dans la manière dont chaque organisation rapporte l'information, particulièrement dans les régions éloignées. Un portrait de l'information communiquée de manière informelle donnerait une idée de l'exactitude des instructions permanentes d'opérations pour la production de rapports d'information, et possiblement, d'en augmenter la modélisation.

Étant donné que le partage d'information entre les organisations n'a rien de banal, nous avons d'abord entrepris une analyse interne des réseaux sociaux au MDN sur le Réseau étendu de la Défense. Mais en raison des changements de priorités apportés par la transformation du CEFC, on a mis fin à Polar Guardian pendant l'étape de la planification de l'acquisition des données, par conséquent, l'analyse des réseaux sociaux n'a pas eu lieu. Le présent document technique présente les leçons retenues dans l'acquisition et la préparation des données sur les courriels ministériels, qui constituent la première étape de l'analyse des réseaux sociaux. Le présent document ne compare pas l'analyse des réseaux sociaux faite à partir de courriels à celle faite à partir de sondages. Les défis (inconnus au départ) de l'analyse à partir des courriels provenaient des contraintes aux plans de l'administration, des politiques, du temps, des coûts et de l'accès à l'expertise technique. Ces défis sont différents de ceux d'une analyse des communications menée dans un environnement d'exploitation commun propre à une étude en particulier; le volume de données est plus grand par exemple. Des questions sont survenues parce que seuls les registres de suivi étaient accessibles (en attendant une approbation juridique, que nous tentions d'obtenir au moment de l'arrêt de Polar Guardian).

Résultats : La quantité et les formes de données d'identification des utilisateurs étaient grandement variables. Nous avons documenté les résultats des efforts à caractériser ces données. Nous avons schématisé les méthodes maison pour l'identification des utilisateurs et indiqué leurs limites, leurs incertitudes et les difficultés. La méthode expéditive d'éliminer toutes les données irréconciliables produirait une analyse d'une exactitude inconnue. Étant donné les contraintes,

[3] Le CEFC a été réorganisé depuis en équipes différentes.

nous avons choisi des outils commerciaux pour générer les données. Nous avons formulé des recommandations pour les analyses des réseaux sociaux futures, y compris de commencer les activités de l'analyse tôt, en raison des longs délais potentiels de résolution.

Comme option de rechange aux solutions maison, nous suggérons une enquête sur les capacités déjà présentes dans le logiciel du serveur de courriels, ainsi qu'un nouvel examen des outils commerciaux qui fournissent des statistiques sommaires (qui utiliseraient un répertoire pour l'identification des utilisateurs, par exemple). Ces outils nécessitent un long délai pour l'obtention des approbations d'accès et pour leur installation. Dans le cas des produits commerciaux, un délai est également requis pour essayer le produit et administrer son achat.

L'accès à des experts de l'administration des systèmes de courriels ou de Microsoft Exchange Server est essentiel, particulièrement s'ils possèdent en plus des connaissances sur l'administration des courriels au sein du MDN. Une expertise de ce genre, en plus d'être nécessaire pour appliquer les solutions aux serveurs de courriels, réduirait le nombre d'éventualités pour lesquelles il faut planifier des solutions. Par ailleurs, ces experts peuvent renseigner sur : (1) la manière dont l'analyse des réseaux sociaux est influencée par les lacunes de la solution dans le domaine de l'identification; (2) les difficultés, les risques et les incertitudes (aux plans technique, administratif et politique); et (3) les ressources et le niveau d'effort requis pour s'adapter aux difficultés et aux incertitudes ou y remédier. Par exemple, en sachant le degré d'empiètement des solutions étudiées sur les restrictions imposées par la sécurité, il est possible d'en tenir compte lors de la planification et de la mise en œuvre des solutions; ainsi, toutes les mesures qui peuvent être prises pour réduire cet empiètement améliorent les chances d'approbation de la solution choisie.

Importance : Le partage d'information entre les organisations entraîne des communications entre différentes organisations dotées de formation et de culture (sociale et d'entreprise) diverses. En comparaison aux communications au sein du MDN, il est raisonnable de s'attendre à une plus grande variation dans le respect des procédures officielles pour signaler promptement de l'information, particulièrement dans le cas des postes éloignés. Une analyse des réseaux sociaux peut indiquer à quel moment et à quel endroit les communications informelles peuvent être importantes aux connaissances de la situation. Les défis découverts, les approches aux solutions possibles et les leçons retenues décrits dans le présent document peuvent informer la planification des analyses de réseaux sociaux futures.

Perspectives : Dans le cas d'une analyse des réseaux sociaux à partir des courriels au sein du MDN, les approches étudiées dans le présent document varient dans le détail. Certaines requièrent un examen approfondi. Par ailleurs, une solution à l'identification des utilisateurs reste encore à mettre en œuvre. Dans le cas d'une analyse des réseaux sociaux entre les organisations, les défis précis demeurent encore largement inconnus, quoique le présent document présente les défis d'ordre général prévus. Les analystes chargés d'une analyse de ce genre devraient travailler avec des experts en la matière pour dresser les défis et les solutions aux plans technique, politique et culturel. Il est attendu qu'une analyse des réseaux sociaux menée au préalable au sein du MDN favoriserait l'adhésion des autres ministères. En outre, bien qu'ils constituent des compromis, il est possible d'utiliser des sondages et des entrevues exclusivement, si la production de données à partir des courriels est au-delà de la portée de l'analyse des réseaux socia ux[4].

[4] En terme d'effort, de ressource et de calendrier.

Table of contents

List of figures

Acknowledgements

The authors would like to thank the following people:

- Dr Phil Farrell for helpful discussions on using Excel® and free database tools for high volume data, for his assistance in connecting with the social network analysts for Multi-National Experiment (MNE) 4, and consultation on selected portions of the text.
- Hannah State-Davey and Mark Round for insights on SNA for MNE 4.
- Mike Manor for the many inquiries he undertook in DIMEI and DND regarding approvals for access to data, and to Microsoft™ support. He was also central to arranging the MessageStats™ demo, which might have gone forward if Polar Guardian continued.
- Donald Messier (Major, retired) for sharing his experiences with email traffic analysis, and inquiries he undertook to obtain details about them
- LCdr Michael Babec for discussions on plausible numbers in estimating requirements for DND internal email and interagency email, and for initiating inquiries into the mailbox users of interest and persons to approach for legal approvals.
- Randy Benoit for discussions that provide a glimpse into the knowledge areas of email administration and directory services.
- Dr Sarah Hill, Dr Katherine Banko, and Brian McKee for informative discussions on the use of rewards for surveys and up-to-date organizational arrangements for oversight of ethical research involving human participants.

This page intentionally left blank.

1 Introduction

Polar Guardian was a project undertaken by the Canadian Forces Experimentation Centre's (CFEC's) Command and Sense (C&S) team[5] to assess and improve situational awareness (SA) in the arctic. Under this project, circa Q3 2006, CFEC's Experimentation Operational Research Team (EXORT) launched a study of the social network of relevant organizations to better understand the flow of information pertaining to SA in the arctic. Due to CFEC's planned transformation into the Canadian Forces Warfare Center (CFWC), this project was put on hold.

This report captures the lessons learned about preparing for such a social network analysis (SNA), to serve as a springboard for future efforts. In particular, this report focuses on an initial email-based phase, which was to be followed by a survey-based phase; these two phases are not treated as separate approaches to be compared. The goal is, as much as possible, to save future executors of SNA from repeating the means-oriented investigations[6] that were performed for email under Polar Guardian, and to guide any investigations with observations from this effort, conclusions, conjectures, and reasoned ramifications. Therefore, this technical note is means-oriented rather than ends-oriented.

Since access to expertise in email administration was somewhat limited, there is a fair amount of reasoned speculation about the requirements and in devising approaches for the data preparation.

Prior to the cessation of Polar Guardian, much of the SNA effort was devoted to finding and liaising with the right people to obtain information required to prepare the data, devising plausible methods to do so in the presence of the challenges and constraints, and vetting candidate tools. After Polar Guardian's cessation, most of the effort was devoted to studying sample email log files and Global Address List (GAL) data to flesh out the ideas for in-house methods for user identification and compilation of data for input into the SNA.

1.1 Motivation for the SNA

This section reviews the history that culminated in the launching of the SNA.

CFEC's C&S team was interested in determining the 'As-Is' capability to maintain SA in the arctic, identifying shortcomings in this capability, and modelling 'To-Be' deployment of assets and doctrine as a remedy. From discussion with C&S, it was determined that the need for SA arises from the following, some of which are mentioned in [1] and [2]:

- The concern has been expressed that many persons/vehicles enter the Canadian North undetected.
- The annual periods during which the Northwest Passage will be navigable are expected to lengthen in coming years.

[5] CFEC has since reorganized into different teams.
[6] As opposed to ends-oriented. Much of the effort was in establishing a means to acquire data for SNA rather than the SNA analysis itself.

- A means is needed to detect terrorism or industrial accidents that result in ecological crisis (e.g. pertaining to shipping, pipelines, or oil pollution), to respond remedially, and to identify and prosecute those responsible.
- Organized crime is attracted to the arctic diamond trade for the purposes of money laundering and manipulation of output diamond quality.
- Drugs and human trafficking are currently commonplace in the North.
- There are territorial disagreements with other nations, e.g. the U.S. and Denmark. Arctic countries are mapping out their continental shelves, since can this can potentially support their offshore territorial claims.
- Search and rescue missions are conducted by the Department of National Defence's (DND's) Joint Task Force North (JTFN). The Royal Canadian Mounted Police (RCMP) has the official responsibility, but often not the capability. Public Safety and Emergency Preparedness Canada (PSEPC) is the main organization for health/safety/emergencies. It is usually supported by DND and the RCMP. DND also supports local authorities, but plays a more active role in arctic regions.

Project Polar Guardian's original emphasis was on modelling the capabilities relevant to SA in order to optimize deployment of surveillance technology and surveillance practices. Information sharing with other government departments (OGDs) and industry was considered an important part of this because of the vast expanse of arctic land, the sparse population, sparse assets, sparse/infrequent surveillance, and the fact that DND is not the only Department that operates in the North.

After investigating possible approaches to modelling the sensor coverage and information sharing, EXORT members suggested separating the modelling for the two aspects. C&S opted to focus first on information sharing between agencies, due to discussions at the Arctic Surveillance Interdepartmental Working Group (ASIWG) about how it would dramatically improve SA in the immediate term. The aim would be to determine whether the right people become aware of relevant sightings and reports under various scenarios.

A major challenge was anticipated in modelling the 'As-Is' information sharing based on formal reporting procedures -- it was not known how rigorously standard operating procedures (SOPs) are followed. The following paragraphs elaborate on two reasons for this. The first is that, in contrast to intra-military information flow, the rigor and uniformity of training in OGDs and industry to follow SOPs for sharing information is unknown, as is the extent to which such formal procedures exist. This variability or uncertainty in following SOPs may be amplified by cultural differences in the arctic, and is compounded by the second reason: understaffing and the possibility of a more casual attitude toward procedures for prompt reporting of information. For these reasons, it would not be unlikely for information to be shared along informal lines of communication.

Regarding the first reason, OGDs and industry operate in different environments and circumstances from the military. Training in steadfastly following doctrinal procedures cannot be expected to be uniform across organizations with cultures that can be very different. To be sure, departure from doctrine is not necessarily bad; in fact, it may be seen as locally adaptive to

circumstances, and potentially an optimization of practice. However, it introduces a large unknown in a model of information sharing based on SOPs.

The second reason for possible disparity between SOPs and actual information dissemination came from EET, whose visit to the Arctic for ASWIG revealed that regulatory offices can be understaffed. This can compromise the rigor with which procedures (particularly administrative paperwork) are followed. Conversely, the remoteness and small community size can result in government offices being in close proximity, thus increasing the likelihood of informal information sharing.

To get a better idea of if and where informal information sharing might play a significant role, a social network analysis (SNA) was proposed. This involves gathering data to show the linkages within a group of offices/people, notionally represented as *nodes* in a *network*, also known as an SNA *graph*[7]. The data can be statistics on volume of communications between nodes, or surveys to educe relationships of various types between nodes. Analysis of the resulting networks can reveal cliques[8] of, and barriers to, information sharing. It can also reveal central nodes that are either bottlenecks or facilitators of information sharing, and individuals that are key to bridging any cliques. The SNA would augment the model of formal reporting pathways, if it did not indicate other approaches to modelling as preferable. The SNA would also be a test of its utility in understanding the flow of information pertaining to arctic SA. Books and articles on SNA are provided in the bibliography, while Annex A provides an introduction to concepts and typical analyses within the context of the popular, free SNA software 'Pajek'.

The SNA would first be performed on DND communications. The results would then be used to encourage the involvement of OGDs and industry.

Analyses that go beyond SNA were also considered, such as tracking the timing of information diffusion[9] e.g. by subject.

1.2 Data Gathering Strategy

Because people do not always behave in the way that they report on a survey, the data for SNA was to be gathered from both communications statistics and surveys. Reference [3] provides a template for building an SNA survey and conducting the analysis. Most of the effort in Polar Guardian's SNA was devoted to arranging the acquisition of email log data and devising approaches for its preparation; the survey would serve as a follow-up phase to the SNA of electronic communications. The aim was to have a social network graph and analysis done in time for the fall ASIWG meeting so that it could be used to educate the participants about SNA. Social network analysts have found that this can be quite engaging for participants; this could used to encourage a high return rate for the survey.

[7] Refer to Annex A for background and example.
[8] "Clique" has a rigorous mathematical definition, but is used here in the general sense.
[9] This was at a discussion level, and the meaning of diffusion was quite open at the time e.g. not only as an email propagated to its recipients, but also who the recipients replied to or forwarded it on to.

While follow-up contact with those being surveyed improves the return rate, it has been found that offering gifts for returned surveys dramatically increases the return rate [3]. The apparent banality of this suggestion belies its importance. A past survey conducted at ASIWG for a different but related purpose had a very low return rate. In conducting a survey, it is generally not acceptable to repeat the survey and ask those surveyed to fill out a form again, simply because the returned data fell short of expectations due to corners cut. As there is no second chance to make a first impression, one way to avoid irreparable shortcomings and maximize the data return is to avoid being too economical with the gift e.g. a nice pen rather than an economy pen. This is especially true for an SNA, due to the interconnected nature of network. Missing links between nodes do not simply give you less data; they can affect the relevant patterns in the network and the conclusions in nontrivial ways. Though the culture of the organization in question will determine how acceptable it is to use gifts as an incentive, the entire cost of conducting the SNA should be kept in mind, as should the price of compromising the accuracy with limited data[10].

The candidate types of information on electronic communication identified for this SNA were *traffic volumes for email and phone calls*. The feasibility of obtaining the data to compile this information, and the adequacy of the data in the records, was to be explored. Email was to be attempted first. As it turns out, this had enough challenges that it became the sole focus.

The duration over which statistics would be compiled would be weeks or months, to establish a steady-state traffic pattern[11]. Statistics would then be compiled for the duration following a significant event of interest, to see whether they differed noticeably from steady-state.

[10] The use of rewards to improve survey return rates is standard practice. Further research is needed, however, to distinguish between its benefits in a public setting versus a corporate setting, and in government specifically. Discussion with SMEs in DND personnel did not reveal a history of rewards being regularly used for surveys within DND specifically, and CF members are not permitted to receive such rewards. If hinterland offices are indeed regularly and severely understaffed, it may be difficult to conceive of a reward that effectively motivates the staff to give priority to a survey. Practices to improve survey return rates require further research (reference [3] briefly mentions alternatives).

[11] This is only to a "first order". There may be an annual or monthly seasonality to the data pattern. The SNA may reveal any such periodicity, as well as the prospects of taking such periodicity into account in "baselining" the traffic.

2 Challenges to Acquisition and Preparation of Data on Email Traffic Volume

This chapter describes the challenges encountered in obtaining email traffic volume data for Polar Guardian. In the course of attempting to resolve these challenges, participants in a past SNA for Multi-National Experiment 4 (MNE 4) [4][5] were consulted. From these discussions, it seemed that many of the challenges arose from the fact that the statistics were being compiled across DND's corporate email rather than for communications in a smaller, more dedicated collaborative environment e.g. such as for an experimental scenario, where the analysts may have more control over the infrastructure or more access to those responsible, files are smaller, there might be only one mail server, fewer administrative and policy barriers, and more direct access to subject matter experts (SMEs). This report will sometimes refer to these challenges as constraints, since a possible way forward may entail working within restrictions (technical or nontechnical) rather than overcoming or removing them. This generally translates into more work, more complicated solutions, or less confidence in the resulting data.

Some of the constraints might not apply to future efforts, depending on the authority behind the request for data, and possibly with much additional time for administrative pursuit. The time frame for Polar Guardian was to obtain example results to bring to the fall ASIWG meeting. Not only would that provide a check of the methodology, but it would also have motivated discussion and hopefully generated buy-in at ASIWG.

The challenges are:

1. **Only the Exchange Server® tracking log files would be provided.** No other data would be generated or provided by the IT administrative personnel.

2. **No pre-processing or filtering of log files by their providers, Director Information Management Engineering and Integration (DIMEI).**

Together with discussions with an executor of a past project involving compilation of email statistics, these lead to the following assumed constraints.

3. **No access to email servers.**

4. **No applications of any kind (commercial or homemade) would be deployed on the Exchange servers**[12] to filter or processing[13] the email traffic volume data, or collect it in any way e.g. by sending distilled statistics to an SQL server.

The above constraints made it necessary to work with the raw tracking log files, which lead to the following challenges.

[12] A specific deployment of Exchange Server® and/or the hosting machine is referred to as an Exchange server (or simply "server").
[13] Any manner of gathering data at the servers, pre-processing the data, or reconstituting it in any way required.

5. **No straightforward scripted/automated access to directory services on the Defence Wide Area Network (DWAN) for email user identification,** if required[14]. Note that this may in fact be possible, with the right expertise, or a similar functionality could be improvised. For example, it was found after Polar Guardian was halted that the DWAN's user lookup database (the Global Address List, or GAL) could be downloaded as a comma-separated-values (CSV) file. This file can be imported into a database that can be locally queried to complete partial identities from the log files. Scripting languages can implement similar functionality by reading the CSV data into a lookup table. However, the speed implications and applications development time would have had to be more fully explored (Section 4.5.1.2).

It was not initially apparent that the DWAN's directory services were needed to identify email sender/recipients. Early communications with DIMEI indicated that email headers were contained in the Exchange Server® tracking log files. In IT, email headers typically refer to specific fields of information, including complete sender and recipient IDs, formatted according to the world-wide standard RFC 2822 (a.k.a. "RFC2822") [6]. When example log files were obtained midway through the data acquisition effort, it became clear that "headers" was interpreted in a much more general sense, and did not contain the required data. The additional step of identifying sender/recipient needed exploration.

6. **Large files handling.** This must be kept in mind when planning how and where to store and process the data to compile the email traffic statistics. Regardless of whether the logs are first filtered before stats are compiled, and whether a database is used, the front end of the data preparation needs to be able to handle large volumes of data. The initial figure provided by DIMEI was 2.5GB of logs per day, and it was envisioned that the SNA would use data over a period of weeks or months. Reduced requirements are estimated in Section 3.

7. **Email server version clarification.** The initial information described the log files as both versions Exchange Server® 5.5 and 2003. Throughout the duration of the SNA effort, it was thought that a major difference between them was that the sender/recipient email address is readily apparent for the latter, but not in many of the transaction records for the former. Accommodations had to be made to process both formats unless/until further information was obtained indicating that only one of the two formats had to be supported. This in fact happened midway through the SNA effort. Discussions with DIMEI personnel indicated that Ottawa servers were the seemingly more cryptic 5.5 version. DIMEI is in the process of migrating to Exchange Server® 2003, however, and the transition would be completed toward the end of 2006.

8. **Question of deducing email sender/recipient ID.** It was not clear what was required to convert the data in the 5.5 logs into unique sender/recipient IDs[15], or whether it was even possible. Conflicting technical information came from different sources (DIMEI, tool vendors).

[14] In this report, a number of approaches to identifying senders/recipients are considered, not all of which require scripted/automated access to directory services. It would be advisable to consult responsible experts to ensure that the manner in which such services are used for an SNA are legal and ethical.
[15] The question of anonymizing user identity data had not yet been broached in this effort, so constraint/challenge#8 refers to identification of a sender/recipient regardless of whether the sender/recipient is anonymized.

9. **Need for information on Exchange Server® 5.5 log files.** The 5.5 log files are in a proprietary Microsoft™ format. Tables are available describing the fields in a general sense, but not sufficiently to decipher the code within the fields. According to Microsoft™ support, Exchange Server® 5.5 is archaic, as is the log file format, and the people familiar with the log file have moved on many years ago. From the tone of the conversation, it seemed that 5.5 was an ad-hoc transitional format in the evolution of the server software.

10. **Legality.** DIMEI had determined that the SNA and/or the acquisition of the email server log files potentially constituted monitoring. They required assurance from legal advisers that the endeavour did not violate ethical or privacy policies.

As an example of past efforts in which not all the above constraints were present (or in which some were dealt with), DIMEI conducted studies in 2004 and 2006 that used email server log files to analyze traffic at a server-to-server level[16]. (In contrast, the SNA requires identification of senders/recipients down to the user level). The goal was to use the traffic data as input to their OpNet network simulation tool. A number of people were involved, including a Microsoft™ Exchange expert. Order-of-magnitude time frames for the data acquisition were as follows.

- The first approach involved several weeks of scripting to extract the email traffic data from the log files.
- The second approach involved two months of approval seeking to deploy custom applications onto the Exchange servers, which extracted distilled email traffic data on a periodic basis and sent it to a separate database.

The second approach avoided the need to process extremely large volumes of email server log data[17], but required significant lead time for approval, as well as Exchange Server® expertise.

[16] Mr. Donald Messier (formerly Major) participated in a feasibility assessment for centralizing the messaging infrastructure. A report "Common E-mail Centralization Study 22 Dec 2004" was produced DIMEI 3-4 for internal DIMEI use. Currently, the DIMEI subgroups have been consolidated into a single DIMEI group. Refer to Annex Section J.1 for current contact details.

[17] It was not established whether the log files used in the DIMEI study were the same as the tracking log files that would be provided for this SNA, nor what order of magnitude were the log file sizes.

3 Data Acquisition and Preparation: Requirements and Considerations

Initially, the plan was to test the use of the SNA on email within CFEC, and then to expand the method to look at traffic between DND mailboxes relevant to arctic SA. In consultation with C&S, it was decided that such mailboxes would be within JTFN and Canada Command[18].

The value of this phased approach became questionable as more information was gathered about the organization of the servers.

1. Pending legal approval, DIMEI was willing to provide any Exchange Server® log files requested. There was no requirement that the value of an SNA be shown on local email traffic before the log files of other servers were provided.

2. From discussion with DIMEI[19], it became clear that the initially estimated 2.5GB of daily log data assumed that EXORT was provided with the tracking logs for all email throughout all of DND. In fact, mailboxes are divided among approximately 100~150 Exchange servers, each handling in the order of 1000 mailboxes and generating a daily log file of in the order of 25MB. From later discussion with CFEC's IT department (Synthetic Environment And Modelling & Simulation Team, or SEAMS Team), it seemed highly likely that users were assigned to servers based roughly on location (geographical and/or on the organizational chart).

With the above knowledge, the focus shifted away from establishing and demonstrating an SNA approach based on the small amount of email within CFEC. The main problem seemed to be determining how many, and which, servers were of interest for the DND traffic. This affects the volume of data that had to be handled, and strongly determines whether any devised approach is practical.

Just as important, however, is ensuring that the method demonstrated on DND email could be feasibly expanded for interagency email. Again, this moves the volume of log data to another level. The issue is compounded by the fact that email will be flowing between multiple domains, the implications of which have not been adequately investigated in this project. With regard to data volume, based on their involvement in ASIWG, C&S estimated the number of relevant agencies to be approximately thirty. This is confirmed in Annex C, where organizations attending ASIWG 2006 were culled from the minutes.

For the purpose of analyzing DND email, it was not known whether JTFN and Canada Command personnel's mailboxes resided together on one server. A server's capacity might allow for such a grouping, but EXORT lacked visibility into the actual groupings. It was prudent, therefore, to make allowances for the division of personnel of interest among more than one server.

[18] To be rigorous, the intra-DND SNA should also include JTFP and JTFA because of their maritime component. However, this was just a trial to establish the SNA process. The final aim was to have an extra-DND SNA that includes OGDs

[19] This includes discussions with Mr. Donald Messier (footnote 16, p. 6) about past studies, as well as with operational personnel in DIMEI about getting server log data.

3.1 Intra-DND Considerations

Together with an estimate of the number of relevant server, the approximate log file size of 25MB/server/day mentioned above can be used for an order-of-magnitude estimate of the volume of the log files to be processed. The estimate of two servers each for JTFN and Canada Command was initially used; later, more margin was given by assuming five to seven servers total. The only way to be sure about the server count was to collect the names of personnel relevant to arctic awareness and look up the servers that they reside on. C&S had initiated an inquiry to get a list of such names, and the information was forthcoming. In the final interagency phase, however, the one-to-two-fold margin for server counts within DND would be of less significance because the possible inclusion of approximately thirty agencies was anticipated.

3.2 Interagency Considerations

For the interagency SNA, the assumption of one server per non-DND agency was considered to be overly optimistic (in terms of simplicity). Since three servers for each of JTFN and Canada Command was assumed, however, and DND likely has more resources than most agencies, the estimate of two servers for each of the thirty agencies seemed to be a reasonable starting point.

3.3 One-to-Many Emails

The edges on an SNA graph (e.g. Figure A-2 and Figure A-3 on p.43) represent person-to-person communications. The obvious way to handle a multi-recipient email with N recipients is to treat it as N "artificial" one-to-one emails. It may be more accurate, however, to count each artificial one-to-one email as less than a real single-recipient email. One need only consider the less relevant broadcast emails that one receives to realize that a single email sent simultaneously to N=20 recipients is unlikely to be worth as much as twenty different emails sent individually to different people, at least in terms of communication that is indicative of a close relationship by which SA information might be informally shared. Intuitively, the broader the audience, the less of a close relationship that the broadcaster has with each individual (at least, as indicated by that particular email). In the extreme case, a message posted to a completely public forum says very little about the relationship between the author and all the possible readers.

The reduced count value of an artificial one-to-one email can be thought of as a weighting factor that attenuates the default count value of 1 for each email message in general. The weighting factor is an open question, but one expects a total weight for all recipients to increase sublinearly[20] with the number of recipients, N. For example, one could weight the total communications N_{Tot} for an N-recipient email as a function of N according to $N_{Tot}(N) = \sqrt{N}$, yielding a per-recipient weight of $1/\sqrt{N}$.

[20] There are rigorous ways to define sublinearity, but here it refers to the behaviour of a single-input, single-output function in the upper-right quadrant of a Cartesian graph (the only region of interest). The goal of the function is to represent diminishing returns, so the slope is always positive, and always diminishes as the independent variable increases.

Since the analyst's judgement is needed on the specific weighting scheme, it is illuminating to consider different weighting examples that can be devised. As an alternative to the unbounded example $N_{Tot}(N) = \sqrt{N}$, consider the case in which one regards emails with recipient lists longer than some threshold (say N_0=30 recipients) as contributing no additional information about relationships between individuals. One possible scheme may be to have the total weight $N_{\mathrm{Tot}}(N)$ for all N recipients defined in so that it is asymptotically bounded by the lesser of $N_{\mathrm{Tot}}(N)=N$ and $N_{\mathrm{Tot}}(N)=N_0$, with a soft transition where the two cross over (Figure 1). The line $N_{\mathrm{Tot}}(N)=N$ is what an N-recipient email would be worth if it was considered the same as N single-recipient emails. The diminishing returns of the actual $N_{\mathrm{Tot}}(N)$ (solid line) expresses the analyst's decision that emails to more than $N=N_0$ begin to enter the regime of mass broadcasts and do not provide much information about individual-to-individual closeness. Annex B provides formulas that can be used for such soft-limiting dependence on N.

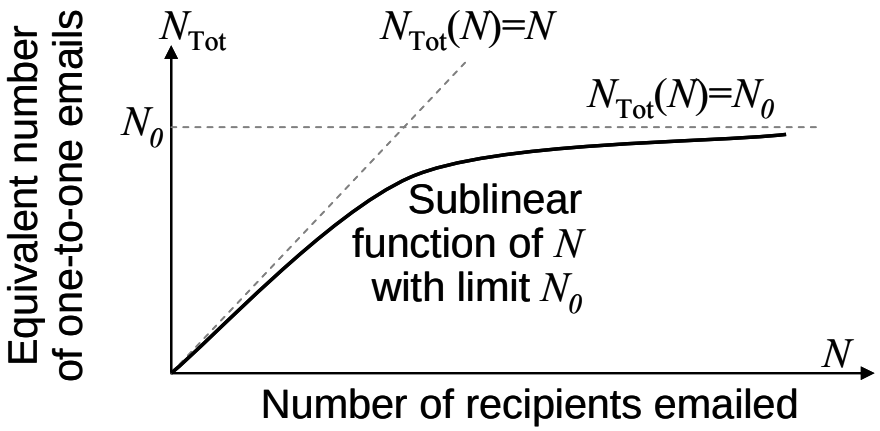

Figure 1: Example of sublinear function $N_{\mathrm{Tot}}(N)$ defined as the lesser of $N_{\mathrm{Tot}}(N)=N$ and $N_{\mathrm{Tot}}(N)=N_0$, with a soft transition where the two cross over

Developing algebraic formulae with which to conveniently weight one-to-many emails raises the question of email discussions among groups of people. Rigorous study of how these discussions manifest themselves in an SNA lies outside the scope of this report, though thoughts are put forth for consideration. First, outright spam is not part of SNA on sharing on SA (though it might be serve other purposes related to IT security). That is, the SNA is envisioned to consider situations in which there is a degree of professionalism and sufficient corporate controls/oversight to prevent flagrant abuse of email, and countermeasures against inadvertent spamming e.g., through computer infection. Hence, one-to-many emails are sent only to groups of people for good

professional reason[21]. These *groups of interest* (GOIs) can be so designated in order to distinguish them from the popular notion of *communities of interest* (COIs); the latter implies a persistent group and doesn't suitably describe ephemeral or one-time issues. Defined in this way, GOIs are a superset of COIs.

Note that sublinearly weighting one-to-many emails does not preclude the manifestation of GOIs in an SNA, since the weighting is merely a preprocessing step applied to the data. Cliques[22] can still be identified within an SNA graph using whatever criteria or means that may be of interest in the absence of sublinear weighting. Sublinear weighting is merely a way to have broadcast emails to (say) N=20 recipients not treated identically as 20 individual emails, since the latter typically (though not always) reflects more time and effort invested by the sender. At the level of a GOI rather than an individual recipient, this reflects the fact that it is quite likely for a broadcast to be of unequal importance (or even relevance) to all N=20 people. If the weighting scheme yields $N_{\mathrm{Tot}}(20)$=13, for example, this still yields a significant portion of the broadcast communication on the part of the sender. Each recipient only registers 0.65 units of communication in this weighting example, but if there is indeed active communications so as to warrant recognition as a GOI, the partial units will accumulate. Furthermore the active disseminators will accrue significant weighting on their outgoing communications to the GOI.

The following requirements follow from the considerations and estimates above regarding data volume, intra-DND email, and interagency email.

3.4 General Requirements and Estimates

In the following discussion, it is assumed that 50 individuals or offices are identified to be of interest of the SNA based on email. The quantity of fifty mailboxes was arrived at based on the aim of overestimating the count beyond any number that can be reasonably expected. This approach was taken because, in consultation with C&S, the actual count of reporting individuals would not likely be obtainable within a short period of time. Gross over-estimation is justified because the potentially prohibitive volume of front-end data is due to the number of servers over which the users of interest (hereafter referred to as *interesting users*) are distributed more than the actual number of interesting users (which is not a very formidable quantity)[23].

1. In the pilot stage, the SNA will include up to fifty mailboxes within DND, distributed across five-to-seven servers, each serving approximately 1000 mailboxes. All servers will be in the same domain, and will consist of Exchange Server® 5.5 servers and Exchange Server® 2003 servers.

2. If it gets to the final stage, the SNA will cover up to 100 mailboxes distributed over approximately sixty servers, not in the same domain, and not in the same corporation[24].

[21] This is admittedly open to interpretation, but again, a rigorous conceptual framework better belongs in a separate study that goes beyond data gathering and preprocessing.
[22] As on page 3, "clique" has a rigorous mathematical definition, but is used here in the general sense.
[23] In other words, the entire (large) log file for a server has to be processed regardless of how many interesting users reside on that server.
[24] The different corporations implies a host of uncertainties. The Exchange Server versions/configuration/operation, users per server, and log file sizes may differ from DND's, if in fact

3. The selected SNA software (Pajek, Annex A) requires data consisting of records, each describing the volume of email between two mailboxes. Hence, the processing of email data should generate data of the following form:

   MailboxA MailboxB Volume_of_Email

   MailboxA MailboxC Volume_of_Email

   ... *etc*. ...

   MailboxB MailboxA Volume_of_Email

   MailboxB MailboxD Volume_of_Email

   ... *etc*. ...

   This requirement on the general form of input data is fairly generic for SNA software.

4. It is *preferable* that the data acquisition method somehow allow for the reduced weighting of multi-recipient emails as per Section 3.3.

5. If commercial software is used to compile the input data for the SNA, it must not have the requirement of being deployed on to the Exchange servers, nor access to Microsoft™ "active directory" (AD) directory service for user identification. The restriction from accessing the AD was to circumvent the long lead time needed to install applications on DWAN. It was discovered after the cessation of Polar Guardian, however, that the GAL can be downloaded as a CSV file. It is not necessarily likely that commercial software will be able to use it in that form, nor is it clear that the GAL data is sufficient to identify users from tracking log data[25]; the GAL's usability should be investigated on a tool by tool basis, and its adequacy should be explored if there is a suitable tool for which it is usable.

3.5 Intra-DND Email : Requirements Estimates and Considerations

In the following, the mailbox counts are order-of-magnitude estimates for an SNA within DND, with the aim of over-estimation, as discussed in Section 3.4.

1. Assume five-to-seven servers hosting mailboxes of interest -- say, six servers, each serving 1000 users and generating a 25MB log file per day, totalling 150MB/day

2. To study email traffic patterns over two weeks (for example) would require processing 2.1GB of log file data.

Exchange Server is even used. The numbers used in this chapters are order-of-magnitude estimates based on DND.

[25] The ethics and legality of combining GAL data with log files would also need to be investigated. This just a specific example of the larger question of whether collecting SNA data in general is ethical and legal, which remains to be investigated. The answer may depend on the details of how the data is gathered and prepared as much as it depends the raw or final SNA data itself. Section 4.6 elaborates on legal compliance for tracking logs specifically.

3. Several tens of mailboxes (up to approximately 100) were expected to be involved in generating/sharing incident reports.

4. For a rough estimate of the number of traffic volume metrics to generate, *M*=50 mailboxes of interest can be paired up in $P = \binom{M}{2} = 1225$ ways. Therefore, up to 1225 metrics of email volume will be generated.

5. For *P* metrics of email volume, if commercial software is used to generate the metrics from the log files, it should ideally not require *P* queries, since the log file data is quite voluminous. For example, if parsing of the log file can occur at 100KB/second, the aforementioned 2.1GB would take approximately six hours. The actual time would depend on the operation of the software, what it does to identify records involving users of interest, and how much of that is done during parsing versus afterward. In any case, the application should be well developed enough that only one pass through the log data is needed, since most of the data will not be relevant.

3.6 Interagency Email: Requirements Estimates and Considerations

In the following, the mailbox counts are order-of-magnitude estimates for mailboxes of interest both within DND and in OGDs, with the aim of over-estimation, as discussed in Section 3.4.

1. Approximately thirty organizations involved in arctic security (Annex C), and which might be included in the SNA.

2. Assume two servers containing mailboxes of interest per organization, yielding sixty servers, each assumed to server 1000 users and generating 25MB/day, totalling 1.5GB/day[26].

3. Over a two-week period, this means 21GB of log files. This doesn't account for the possibility that traffic is less on weekends, but again, this is an order-of-magnitude estimation with a leaning toward over-estimation to provide margin.

4. Several tens of mailboxes (up to approximately 100) are expected to be involved in generating/sharing incident reports.

5. Estimate: *M*=100 mailboxes requires up to $P = \binom{M}{2} = 4950$ metrics of email volume.

6. Even more critically than for intra-DND email, for *P* metrics of email volume, the tool used for data processing should not require *P* queries of the log file data, since the data set is extremely large. (If queries are done on a database of records pertaining only to mailboxes of interest, however, this requirement may not be as important).

7. The servers that generate the log files are not on the same domain.

[26] Order-of-magnitude estimates based on DND.

8. If commercial software is used to compile email volume metrics from tracking logs, it should be able to maintain distinct mailbox identities, even though the servers generating the log files are on different domains.[27]

[27] If this is not technically possible, then alternatives to tracking logs need to be investigated for generating SNA data. This of course has implications for the feasibility of the interagency email SNA as a whole.

4 Addressing Requirements Estimates and Considerations

This section documents:

1. The selection of SNA software
2. Vetting of commercial software for data gathering and preparation
3. Aspects of previous attempts within DND to study email traffic volume, which could inform the method of preparing input data for the SNA
4. Selection of a scripting language for in-house approaches to data preparation
5. Knowledge gleaned from examining the example tracking log files, on which approaches can be based for user identification
6. Legal issues and likelihood of their resolution
7. The suitable computing network for data preparation and the SNA

These points are discussed in sections 4.1 to 4.7, respectively.

4.1 SNA Software

Three packages that are among the most popular SNA software available are UCINET, Pajek, and NetDraw. To circumvent the administration of purchasing software, EXORT opted to use freeware, which excludes UCINET. Pajek and NetDraw both seemed well documented, so the choice of which to start with was somewhat arbitrary. Pajek (Annex A) had the benefit of having an SNA textbook based on it [7], however, so it was chosen as the package with which to start exploration of SNA.

4.2 Data Acquisition Software

The initial plan was to write scripts to preprocess the log file data for input into Pajek. DIMEI's advocacy for Quest®'s MessageStats™ tool prompted the exploration of that avenue as a possibly quicker solution. MNE[28] 4's use of Excel® for their SNA prompted the examination of Excel®-based solutions. Discussion with SEAMS on SQL Server requirements generated a series of potential commercial alternatives to MessageStats™ and/or SQL Server. In researching those options, further commercial alternatives were encountered. These investigations are detailed in Annex D.

[28] Multi-National Experiment.

After the above investigations, the initial plan to process the data through scripting seemed to be the most direct, simple, and realizable. It had the least amount of uncontrolled dependencies that could render the solution unusable e.g. dependence on software vendors, purchasing, authorization for installation, and installation. Since EXORT would only have access to the tracking log files, however, the issue of unambiguously identifying sender and recipient (with confidence) needs resolving for any of the approaches to be workable.

The following is a summary of the commercial software approaches in Annex D, investigated circa September 2006. The accuracy of information on commercial products is limited to the accuracy with which the information was provided in consultations with the vendors.

The use of Excel to convert the raw log data into an Access database was limited by Excel's low maximum record count (65,536), the nontabular nature of the log entries, and inadequate user identification data. The same limitations apply to Excel's "PivotTable" [8] feature, which was used to compile SNA input data in MNE 4.

The raw log data can be reduced in volume by pre-filtering, and made tabular by converting multi-recipient email entries to one-to-one equivalents. However, this pre-processing phase can also be made to compile the SNA input data (Section 4.4), thus obviating the need for Excel, PivotTable, and Access. Unfortunately, it does not solve the problem of obscure user identification data.

Of the remaining potential commercial solutions, the most investigated was Quest's MessageStats, due to advocacy by DIMEI. Despite ample communications, however, demonstration of its functional suitability (in the absence of the Active Directory) and affordability was still forthcoming, and it was not clear whether EXORT would be able to meet requirements regarding database size and software.

The remaining commercial options were considered less promising for one or more of the following reasons, in order of decreasing insurmountability[29].

- Lack of response
- Uncertain functional suitability (proper transformation of log data to SNA input data)
- Access to Exchange servers required
- Access to Active Directory required
- Requirement for SQL Server
- Cost

4.3 Leveraging DIMEI's Experience

DIMEI[30], which conducted the email traffic studies described at the end of Section 2, raised the possibility of leveraging this work in gathering data for server-to-server traffic. After further

[29] As subjectively viewed by the primary analyst of this SNA.
[30] Discussion with Mr. Messier (Footnote 16, page 6).

investigation, however, it was deemed unsuitable for the SNA, which requires data at the resolution of individual users. The possibility was also raised of guidance from DIMEI in navigating and expediting the process of approval in the (unlikely) case that EXORT ended up deploying tools/applications onto the Exchange servers.

4.4 Scripted Data Preparation With Perl

If it is assumed that tracking log files are the only starting point, it seems in hindsight that time spent exploring potentially more elegant commercial solutions to data preparation might have been better spent on the initial simple approach of scripting the functionality, low level though it may be. Two of the most popular languages meant specifically for data processing are Perl and Python. Interpreters for both are free. Perl is the more mature of the two, and extensive knowledge and ramp-up material exists in the public domain. Since the primary analyst of this SNA had previous exposure to Perl, it was chosen as the language with which to process log files into Pajek input data. Annex E contains reference material for Perl.

4.5 Deciphering Exchange Server® Tracking Log Files

As mentioned, deciphering the user IDs in the tracking log files is necessary because of lack of access to the Exchange servers (the Exchange Server® applications running on the host machines, and associated report generation capabilities), the inability to deploy custom applications on the host machines to collect the required data in a more easily used form, and the near term barriers to installing commercial applications on the DWAN to access the AD. A completely reliable method for discerning the user IDs from the logs has not yet been devised. The information in this section was captured from efforts taken after the cessation of Polar Guardian. For future SNAs in which the use of tracking logs is explored, this starting point circumvents the bottom-up discovery that has already been paid for in Polar Guardian project time. Outstanding issues of uncertainty include:

1. The lack of sufficiently comprehensive, publicly accessible documentation on the log files

2. The need for, and potential impediments to, high-speed identification of users

3. The question of suitability and adequacy of the tracking log files, which appeared to be plagued with
   - A plethora of types of potentially identifying data, with varying degrees of potential ambiguity
   - Unpredictability in the presence of these various types
   - Discrepancies in their format

Most of the effort was devoted to the study of the example Server 5.5 log, since that was the format of the logs at the time. The transition to Server 2003 was imminent, however. Despite that, information pertaining to both formats was kept because the mail servers in OGDs are not known.

One should also keep in mind the remote possibility that some OGDs do not use Microsoft™ Exchange Server® (of any version) at all, since any log files for such servers will have yet another format. It is not known whether the identification data in non-Microsoft™ server logs is as unpredictable as in Exchange Server® 5.5 tracking logs.

The arcane nature of the identity information in DND's tracking log file could very well depend on the viewer's degree of expertise in email protocols. It is conceivable that someone with extensive training in email administration would find the data recognizable without more in-depth documentation. It is not known how common such expertise is. For Polar Guardian, exploration of this possibility ended up in a referral to Microsoft™ by DIMEI. The ensuing discussion led to the assessment that home-grown deciphering of the tracking logs was not advisable. Since it is the tracking logs that would be conditionally provided, however, forensics was performed on example log files to generate potential approaches to determine identity.

4.5.1 Exchange Server® 5.5 Tracking Logs

Exchange Server® 5.5 log files and events are described in Annex F. According to Microsoft™ support, there is no further, more elaborative documentation. As per DIMEI's warning, the log files were indeed found to be cryptic; it was unclear for some time whether sender/recipient IDs could be deduced from the records. Anonymized examples from the tracking logs are contained in Annex H. The tracking logs became clearer after the cessation of Polar Guardian; contact was finally made with the appropriate personnel at Microsoft™ support, and the example log files were studied extensively.

4.5.1.1 Information from Microsoft™

As mentioned, Exchange Server® 5.5 is quite old. Microsoft™ managed to contact a former engineer who had exposure to it and was able to provide the following opinions based on anecdote. Note that the level of expertise in email protocols of the Microsoft™ participants in the associated discussions is not known.

1. No expertise exists in Microsoft™ to convert the log file data into sender/recipient email addresses, as the technical people working on that transitional format have moved on circa 2000

2. Had they been around, the solution would likely be quite involved

3. There was scepticism about the prospect of connecting to directory services to resolve the sender/recipient addresses. It wasn't clear whether it was the actual connecting to the service which was difficult, whether the tracking log was deemed to have insufficient information to resolve the sender/recipient identities using the directory services, or whether the directory service was the GAL or AD.

4. Approximately 80% of the transactions (i.e. records) in the logs did not represent delivered email; they were intermediate hops from server-to-server, a series of which make up the end-to-end delivery of email. (This simply meant that there are, on average, four hops in the delivery of each email).

5. The actual delivery of an email could be recognized by the field identifying the event associated with the transaction. The event should be something to the effect "message delivered to local store".

Subsequent consultation of the event definitions (Annex F) reveals the closest event description to "message delivered to local store" is "The MTA completed delivery of a message to local recipients (usually through the information store)" (event#9). A study of the sample log file confirmed that such events comprise 24% of the records. The complementary event in Annex F seems to be #4 ("Message submission"), which comprises 18% of the records. Both 18% and 24% are in line with the estimate that 80% of a server's traffic merely uses the server as an intermediate hopping point. The fact that outgoing messages are fewer than incoming messages seems reasonable, since a multi-recipient message will count once as outgoing, but many times as incoming[31].

Since a small minority of the records are relevant (non-hop events, respectively numbered "9" and "4" for receipt and dispatch of messages), the prospects of reading all the relevant data into Excel® and using PivotTable® (Section D.2) to generate SNA input data are improved, though the caveats still apply.

The prospects of using any of the data preparation approaches based on tracking logs initially seemed to be rendered moot, however, by the above assessment that sender/recipient ID cannot be practically reverse-engineered from the 5.5 tracking log files. Fortunately, scrutiny of the sample log file and some research into standards led to some possibilities for identification[32], as discussed in the following sections.

4.5.1.2 General Observations on Identification Data

Annex I contains the details of the observations on identification data.

The prospects of identification are further improved by the availability of GAL export data, discovered after cessation of Polar Guardian. From speaking with IT personnel, it was found that the GAL data is available as a file on DWAN at http://img.mil.ca/natsvcs/cfnoc_hd/DEMS/f_downloads_e.asp. This information improves the feasibility of writing custom applications to perform identity translation/completion, though as this section discusses, challenges still remain.

Many of the identity records in DWAN's GAL database contain data items vaguely resembling the cryptic subfields in some of the 5.5 log fields. The GAL database lists this data for each user as "X.400" data, which is an alternative standard to SMTP (Simple Mail Transfer Protocol), the protocol for delivering RFC2822 messages. X.400 has not come into the mainstream, but is apparently used in the military, intelligence, and aviation.

In the example 5.5 log file provided by DIMEI, the X.400-like data occurs not necessarily in the fields for sender or recipient, but seemingly always in the "Message ID" field. Therefore, it does not directly resolve the identification problem. In conventional SMTP, however, the message ID

[31] Hypothesized rationale.
[32] Feasibility and level of effort need estimating.

typically contains information that can help identify the source of email; further consultation with SMEs is needed to determine if the message IDs in the log files can help resolve the identification difficulties described in this report.

Each of the sender and recipient fields contains an unpredictable combination of data items, presumably about the same person. It remains to be determined whether one person can be identified with different combinations of data types in different records. When present, certain data items are preferable to others. Ranked in terms of least to most ambiguity in identification, they are:

1. SMTP address

2. X.400 data, since it may include middle name initials (in contrast to X.500 data, described next)

3. X.500 data, which may contain first and last name only, and possibly not even that. X.500 is a standard for directory services and supports X.400. In the tracking log data, X.500 data is often followed by three-digit numbers; these turn out to be not unique to individual users, and hence do help resolve user IDs in an obvious way.

These data types in the tracking logs were discerned by comparing the log data with both the downloaded GAL database, and the GAL as accessed from Outlook. At the time of writing, these two methods of accessing GAL information have yielded non-identical information, irrespective of whether one downloads the GAL database for Exchange Server® 5.5 or 2003. When accessed from Outlook, the GAL contains both X.400 and X.500 information for the users; in contrast, the downloaded databases do not appear to contain X.500 information.

If GAL data *must* be used to identify a user based on X.500 data, therefore, it would be necessary to be on the DWAN in order access the online directory rather than use the downloaded GAL e.g. it may be necessary to investigate whether the X.500 data can be obtained via the AD rather than the downloadable GAL databases, as well as methods or tools for doing this. Having such access on the DWAN, however, goes beyond the constraints being observed in this SNA.

4.5.1.3 Observations on Non-Intermediate Hops

Since the log transactions corresponding to intermediate hops contribute only noise to the identification process, the records for non-intermediate hops were examined to see if the data was more identification-friendly. These transactions are *outgoing* and *incoming* emails; from Section 4.5.1.1, these are likely to be event numbers four and nine, respectively. To distinguish them from intermediate hops, the understanding in this terminology is that they are outgoing from, and incoming to, *mailboxes* rather than *servers*.

In principle, end-to-end email within DND can be culled from the logs by taking only the outgoing transactions *or* the incoming transactions. Email with organizations outside of DND, however, require that both incoming and outgoing emails be examined. Under such a counting scheme, internal email will be double-counted; conceptually, however, correcting this is expected to be straightforward.

As detailed in Annex I, records for outgoing messages seem to always have sender first and last names. Though it doesn't always identify a user uniquely, the potential ambiguity is limited. In contrast, the recipient field often contains cryptic information. The reverse is true for incoming messages. More research or consultation with subject matter experts (SMEs) is required to determine whether identification can be made for all the cryptic identifiers.

An interesting, though theoretical, possibility for avoiding the irresolvable identities arises from the fact that the sender has good identification data in outgoing email, while the recipient has good identification data in incoming email. Their limited potential for ambiguity will be ignored for the purpose of describing this scheme. The possibility for identification is premised on the assumption (based on the SMTP world) that the message ID remains the same regardless of the path taken for message delivery. Outgoing messages can be matched up with incoming messages of the same message ID, which allows the identification of both sender and recipient. The message then contributes to the message count for the SNA link between their corresponding mailboxes. There are caveats:

- It is theoretical because the tractability of matching corresponding incoming and outgoing email needs to be investigated, and can be a potentially involved capability to develop in-house
- Whether the message ID is constant throughout the delivery process needs confirmation
- Both outgoing and incoming records are needed
- Though some of the identification is good (sender in outgoing email, recipient in incoming email), caveats in Section 4.5.1.2 may still apply.

4.5.2 Caveats in Using GAL Information to Discern User IDs from Exchange Server® 5.5 Tracking Logs

In addition to those described in Section 4.5.1.1, this section discusses caveats that should be considered in devising an identification scheme that uses the GAL, including:

- Anticipated high speed requirement for GAL-assisted translation
- Ambiguity in using first/last name of X.400/500 data in the tracking log files
- Discrepancy in GAL information
- As an alternative to the GAL, manually drawing up a name table to resolve user IDs
- Considerations due to the many forms of user ID in the tracking logs

If it turns out that GAL-assisted user identification is required to filter away transactions not involving users of interest, then the large volume of raw log data requires that the identity resolution be very fast. This would depend, for example, on the implementation details of the lookup table created from the downloaded GAL database.

From the X.400/500-like data examples in Annex I, it seems possible that a complete identification can be made from the surname and given name. The greatest drawback to this is that such information is not present in all transaction records. Furthermore, users of the GAL will realize that first and last names often do not uniquely identify a person. This is the

aforementioned limited ambiguity of identification using first and last names, and can be mitigated by selecting only servers hosting users in Canada Command and JTFN. It is also further mitigated in the case where a transaction record has X.400 data that contains the user's initials[33]. There did not appear to be any middle name initials for the X.500 data with which to mitigate ambiguity.

For the tracking log records in which they are available, use of names in the X.400 data is complicated by the fact the X.400 data is part of the aforementioned discrepancy between the downloaded GAL database and the GAL info as accessed from Outlook. For example, it was observed that compound names can be hyphenated in the latter, but simply attached together in the former. The reason for this discrepancy needs to be determined before there can be confidence about any lookup table built to resolve IDs. If the lookup table is built from the downloaded GAL data, it would be far more convenient if the X.400 names in the tracking logs matched the downloaded GAL data. This was indeed observed, but only from a cursory examination.

Even if some tracking log identities match the GAL as accessed from Outlook, however, just knowing that the discrepancy is *consistent* would allow for manually changing the lookup table to compensate. For this to be done, of course, all such discrepancies need to be identified. The level of effort to do this needs to be scoped out.

The above example of manually tailoring the lookup table of first/last names can be taken to the extreme; if the user names in the logs are always the same, it is possible to forego basing the lookup table on the downloaded GAL database by manually drawing up the translation data based on the names in the logs. This is only tractable, of course, if the number of users of interest is small enough. Again, this will only help for the records in which the *proper* X.400/500 data is available i.e. those that include first and last names.

Despite these workarounds to the specific, observed examples of problematic data, a major challenge is the fact that user ID fields contain a mixture of different types of information that is highly varied across transaction records. This complicates the devising of an identification scheme. To account for the various possibilities, it is likely that the final scheme will take longer to implement and be more computationally demanding i.e. run slower.

The absence of X.500-like data in the downloaded GAL database would render that information in the tracking logs unusable for GAL-based identification. For records containing X.500 data, the only way to use the data would be to base the identification on first and last name only, using a custom lookup table built for such, as described above. Such a solution also inherits the potential ambiguity in identification based on first and last name.

A solution to the construction of a lookup table is possible if the different ways of identifying a user are predictable and limited. A simple first approach might be to have multiple entries of the lookup table translate the data into a common user. The as-yet unsolved challenge is ensuring

[33] For the purpose of this report, the user account is taken to be the most unambiguous identification. It is far from impossible in an organization the size of DND/CF to have people with same first and last names, and possibly even the same middle initial (though less likely, of course).

that all possible variations of a user are identified. They will then occupy separate rows in the lookup table.

The forms of "identification data" for which accommodations cannot be made, however, even in concept, are those examples in Annex I that have no known association with a user, either through the presence of first and last names, SMTP address, or some discernible aliasing.

4.5.3 Exchange Server® 2003 Tracking Logs

If SNA is performed on Exchange email in the future, it is quite likely that the tracking logs will be in Exchange Server 2003 format. This is described in Annex G. With the rare exception, the events for Exchange Server 2003 are defined almost identically to those in Exchange Server® 5.5 over the range of event numbers for 5.5. It was noted, however, that the 2003 events have a series of new events *above* the range of the 5.5 events, and they are almost exclusively described as SMTP related.

As mentioned, initial examination of a sample log file from DIMEI seemed to indicate no complications in identifying sender and recipient; all IDs seemed to be SMTP addresses. However, the log file was quite large (85MB, over 300K transactions), and a deeper examination shows that the sender/recipient IDs suffer the same kind of problem as the 5.5 log, though to a lesser degree. That is, some of the IDs are not recognizable email addresses, though most are SMTP format. Many such records appear to be X.500 data, but not all.

It is possible that the differing degrees of missing email addresses between the example logs for Exchange Server® 5.5 and 2003 are due to different roles in the servers that generated them, and are not merely due to the version of Exchange Server® software. This suspicion arises from studying the ranges and distributions of the event numbers. As shown in Annex F, the 5.5 log files have event numbers discontinuously defined in the two ranges 0-52 and 1000-1018. The descriptions for the high range of events seem to deal almost exclusively with SMTP and nonlocal delivery, in apparent contrast with events in the low range. Only 14% of the records in the 5.5 example log have events in the high range, and *all* such events are numbered exactly 1000 (which indicates that sender and recipient occupy the same server). This can be contrasted with the example 2003 log, in which *all* events are well distributed *above* 1018; this range is not defined for 5.5 logs and deals almost exclusively with SMTP. Because of the different types of events, it is possible that the 2003 log file came from a server playing a more specialized role.

Caution is warranted, therefore, in assuming how representative the example log files are. If the 2003 example log came from a gateway server that connects DND to the outside world, for example, a logical question is whether the log can provide any insights at all into the decipherability of logs for user-hosting servers. In fact, answers are needed to the questions of whether mailbox hosts and gateways require separate servers, and what other server roles may and may not share the same server as mailbox hosts, and of those that may, whether any transactions are logged that are irrelevant to identifying end-to-end email volume. Such log entries would *appear* to obfuscate the log files but would actually be ignored in front-end filtering. They do not add to complications in user identification for end-to-end email, but do increase the volume of front-end data.

4.6 Legal compliance

Prior to cessation of Polar Guardian, C&S was in the process of determining the appropriate persons with whom to consult for legal approval to receive data on DND email. Information about the appropriate persons was still forthcoming, but C&S felt that approval would have been readily obtained, if:

1. EXORT and/or C&S can give assurances to DIMEI that access to, and use of, the data in their possession would be controlled e.g. by stipulating which team members will be allowed access to the data, under what conditions, and how those individuals will handle the data to prevent inadvertent proliferation, including but not limited to the media, networks, and hosts which can carry the data.

2. There would be commitments to safeguard against abuse of privacy.

Analysts might also inquire with DIMEI (footnote 16, p. 6) for guidance in navigating the issues of approval, since his group must have had to deal with them in obtaining the logs for his study. There are various types of server logs, however, and it should be confirmed that the logs used by DIMEI are the same as the tracking logs that DIMEI has made conditionally available for the SNA.

Since DIMEI will conditionally provide the tracking log files with no pre-processing or filtering, the assurances of due diligence and legal approval should cover more than just the use of the data for the SNA. It must cover the fact that all of the data in the raw log files will be delivered, even though most of it will be ignored. This is one of the reasons why obtaining more explanatory documentation of the log files was so important; it allows a clear description of what the data means, and what can be inferred from it.

Aside from DND email, the prospects of obtaining necessary approvals from OGDs are another matter still to be investigated. It is not expected to be straightforward.

4.7 Data Preparation Outside of DWAN

The DWAN policies against installation of non-standard software are quite restrictive. Since the options for meeting the data preparation requirements were speculative, it would necessary to use a more flexible computing environment such as DRENET so that the required applications can be installed to accommodate the possible contingencies. Local admin rights to install software and computing/analysis environments on DRENET are sometimes granted on a case-by-case basis, depending on the need.

An initial concern was the possibility of restrictions on where the acquired log files can be hosted. If they had to remain on DWAN, the ability to implement the data preparation approaches would be severely limited. DIMEI has confirmed that analyzing the data on DRENET is not a problem, however, as long as legal approval to access the data is obtained.

Aside from DRENET, other options for preparing the data and/or the subsequent SNA on resulting traffic metrics can be performed on a stand-alone system. In future SNAs that EXORT

may conduct, the work can also be done on the cubicle-area network that was being planned by EXORT at the time of writing.

5 Lessons Learned

Section 3 presented requirements, estimates, and considerations that arise from considering likely numbers of users, servers, and organizations for the SNA, as well as the user/server organization within DND. Section 4 vetted tools and approaches for preparing input data for the SNA, and captured characteristics and implications in the unfavourable case where tracking logs are used to generate this data. Issues on legal approval and migration of data to a hospitable analysis environment were touched upon.

In addition to overall lessons, this section discusses:

- Activities that should be initiated as early as possible in an SNA study because of potentially long resolution times
- Caveats related to commercial solutions for data preparation
- The need for access to subject matter expertise in email administration
- Anticipations for an interagency SNA
- Suggestions for a follow-up survey

The lessons and recommendations extrapolate from the experience in Polar Guardian's SNA and can therefore be speculative in nature. This is particularly true in view of the fact that many of the issues result from incomplete information, and the planning of courses of action for the contingencies. The questions raised might help direct future efforts. Much of the discussion revolves around the need for subject matter expertise in assessing options for data preparation. Within the context of a future effort (resources, time frame, authority, constraints), a SME[34] can help determine which issues are resolvable trivially or with a reasonable amount of research. Some issues might not be resolvable within the limits of the project or might not be relevant in the circumstances of the study. Some speculations may also be off-base, but awareness of their possibility could result in more incisive discussions with solution vendors or in-house personnel.

5.1 Reconsidering the Constraints, Time Frame, and Approaches

- When studying electronic communication logs (as opposed to using surveys), SNA requires a lot of high-volume data processing. Data source collection planning must be thoroughly considered in advance.

  This is not merely an observation from the Polar Guardian effort. It corroborates with a discussion with DIMEI (footnote 16, p. 6) about his server-to-server traffic study, and with communication with the social network analyst for MNE 4.

- In the context of thoroughly considering the data collection, the whole approach of using only the tracking log files for email traffic volume needs re-examination.

[34] SME: subject matter expert.

The use of tracking logs was based on the premise that email header information was captured in those logs. When this turned out not to be the case, the prospects of user identification became extremely unclear. Most of the effort was then devoted to exploring possible methods for identification of users, as consistently and reliably as possible, but in a short time frame. This precluded installing home-grown or commercial applications on DWAN or the mail servers, since it incurs the risk and delay of authorization, and cost in the case of commercial solutions. Challenges and uncertainties remain.

- Considering the feasibility challenges that arise from the above constraints, it is highly advisable to expand the time horizon for data preparation and further investigate options that involve data collection/preparation done at least partly on the mail servers and/or by applications on DWAN machines that can access network services such as the AD.
- Though it costs time and may be expensive to install applications on the mail servers or DWAN, and approval may be uncertain, it is in no way clear that less time would be needed to solve the user identification problem with the tracking logs in isolation (with the possible help of downloaded GAL data). For the latter, reliable and consistent identification may even be infeasible, as indicated by discussions with Microsoft™ and two solution vendors.
- How the data gathering and preparation might be implemented on the mail servers requires further investigation.
    - Discussion with DIMEI (footnote 16, p. 6) may provide a starting point.
    - The investigation should include finding out what the tools that come with Exchange Server® are capable of, what suitable software may reside on the server machine (if that is possible) as a peer application with the server, and whether suitable applications can be written on top of the functionality provided by server tools.
- Software to be installed on DWAN to access the AD could face smaller approval barriers and less delay than implementing solutions on the server, and should be reconsidered.

  Several commercial packages were rejected because they needed such access to the AD to identify users in the log records. With a longer time horizon, it is recommended that social network analysts re-visit these options.
- In an SNA of electronic communication where user identification data is *not* an issue, if the data set is small enough, the use of Excel®'s PivotTable® feature can be used for the task of data preparation with minimal fuss.

5.2 Things to Start Addressing Early in a Study

- In an SNA of electronic communication, insist on getting sample data up-front to check the usability of the data and the level of effort required to make it usable.

  As illustrated in Polar Guardian, the request for information can sometimes be interpreted in a manner that was not anticipated, and the reality is that the usability is much less than one may have been led to believe. If making the data usable takes much more effort and involves much more uncertainty than originally envisioned, then the SNA itself, the approach to gathering data, and/or the timeframe needs re-examination.

- Get the approval for accessing the data from the responsible authority, in case any time-consuming preconditions need satisfying.
- Until data is actually received, even expressed approval should be viewed merely as intent. Nontrivial conditions may be imposed at any time. For SNA in particular, the use of data on interpersonal interactions inevitably raises issues of security and privacy. In the case of Polar Guardian, the initial approval for data was later prefaced with a requirement for legal assurance to address those issues.
- Initiate inquiry into legal approval for obtaining data on electronic communications early in the study, including inquiry concerning what legal authorities need(s) to be contacted. It could take some time. Due to the sensitive nature of email data, future SNAs based on email log data would almost certainly elicit similar concerns (as opposed to SNA of communications in experiment-specific common operating environments). Beyond legality, consultation with Director Military Personnel Operational Research and Analysis (DMPORA) is needed to clarify the restrictions and obligations arising from ethical considerations[35].
- The identity of those people whose communications are relevant to the SNA are required ahead of time to determine the number of servers on which they reside, the expected volume of data, and possibly even the cost of any commercial software licenses. Getting these names depends on responses from different people, and can take time.

5.3 Caveats for Commercial Packages for Data Preparation

- Until one actually sees a commercial solution perform the preparation of the required data, the stated or implied suitability of a tool or approach cannot be taken for granted.
    - An analyst can describe functional requirements to potential tool vendors, but they are not always carefully read, or the suitability of a candidate tool may be casually posited.
    - In the case of MessageStats™, confirmation of cost and functional suitability was forthcoming from the vendor for quite some time, as was the demo, but did not materialize
    - Additional requirements, such as the need for SQL Server in several cases, could delay or stall a solution. SQL Server in particular seems to be a commonly encountered requirement. Its absence in an organization that deals with generating and analyzing data might not indicate a shortcoming in the commercial solution so much as it indicates a capability requirement in the performing organization. The IT department for CFEC (SEAMS) has expressed

[35] Within DND, a former incarnation of an oversight body for ethical research involving human subjects is described in [9]. This is a research ethics board involving Director Military Personnel Strategy (DMP Strat) and Centre for Operational Research and Analysis (CORA). This role now falls to DMPORA, which also coordinates surveys across DND to avoid over-surveying segments of personnel. Such oversight would be essential for a survey-based SNA.

recognition of a vendor-agnostic need for this capability[36], beyond the free (and more limited) alternatives.

- Despite the potential advantage of commercial tools, there are components of time overhead and risk to be taken into consideration, as per the following examples.
  - The search for candidates and the exploration and confirmation of their suitability can take quite a bit of time.
  - The approval for their purchase and installation introduces further risk and delay; this also applies to any prerequisite software not already available.
  - If the software needs to be installed on DWAN specifically e.g. to access the AD, the approval for installation introduces more risk and delay than if it can be installed and used on a less restrictive environment for research and data analysis, such as DRENET. (For the SNA, however, the AD of interest is on DWAN).
  - Any deployment of data gathering applications on the Exchange servers themselves introduces significant risk and delay. Depending on the capabilities that come with Exchange Server®, however, it might be possible to get data in suitable form without purchasing commercial reporting tools.

  These factors need to be investigated and weighed in determining which tools will serve as alternatives to a complete in-house scripting approach.

- The robustness of commercial solutions needs to be assessed.

  Some of the relevant records in the mail server logs look like they wouldn't be amenable to any kind of identification. If deep knowledge about the server logs is as rare as Microsoft™ suggests, then it is conceivable that such records are simply ignored by commercial packages. This begs the question of whether they are any better than an imperfect home-grown solution that potentially uses GAL data rather than the AD, *and* ignores the irresolvable identities.

5.4 General Access to Expertise in Email Administration, Exchange Server®, and Directory Services

- For future planning purposes, note that end-user access to Microsoft™ support is controlled relatively carefully within DND and may take time to initiate.

  A series of communications within DND can be expected before making contact with Microsoft™, after which a series of communications can be expected in order to reach the person with the relevant expertise. Before relying on this path, be aware that it could take weeks. It can be necessary, however, if there is lack of in-house expertise, lack of time on the part of any in-house experts, or understandably, lack of priority given to research projects relative to operational needs.

[36] While need for the generic capability was recognized, there was care to articulate this need in a manner that was independent of a particular tool or product so as not to unduly bias the future choice of a solution.

- It is highly beneficial to have involved in the SNA a SME in Exchange Server® and email protocols, *and* to have access to this person[37]. It would be even better for responsibility to the SNA to be part of the work plan of a person within the DIMEI hierarchy. The impact that this would have on an email-based SNA is highlighted by the following circumstances.
- In Polar Guardian, key information was encountered in a piecemeal and random manner, or encountered in web searches to decipher the log file

  Examples include information about email protocols, directory services (such as that for user identification), Exchange Server®, and the likely organization of users among servers. Having direct access to a SME would greatly accelerate this indirect method of bottom-up knowledge building. Without such an expert, much of the information in this report would not have been acquired were it not for fortuitous networking and the taking of every opportunity to direct conversation toward the topic of email administration.
- In-house personnel who were able to contribute knowledge had operational responsibilities that precluded the necessary involvement in the planning of data collection, both in terms of degree and timeliness.

  Therefore, in addition to the networking overhead to find the right persons to connect with, there is also overhead in continually maintaining tactful communications until their operational priorities were adequately dealt with to permit response on the fragments of information being sought.
- Access to required knowledge can be hampered because direct communication with the most relevant in-house experts rather than the formal point of contact (POC), without formally established channels of communication, is not always appropriate.

  Key information about the server versions in DND and their transition in the near future, for example, would never have been encountered were it not for a conversation with a contractor in DIMEI; such discussions constitute lateral communication. Basically, "legitimate" and approved channels of communication would not have allowed the gleaning of important facts because one has to know what to ask for before fielding the right questions through the formal POC, higher up in the chain of command. The extra intervening links also increase delay, with multiple follow-ups, since the POC may be more senior, quite busy, and have operational crises to solve. Timely confirmation of information from the SME is not possible; this leads to nontrivial miscommunications, such as positing that the tracking logs contained email headers.
- Much of the required email knowledge and networks is not necessarily cutting edge, but is in a highly specialized domain and sometimes difficult to find, access, or utilize.

  Documentation is sometimes not obtainable publicly, or requires augmentation with expertise and experience to be quickly meaningful to those outside of the field e.g. standards for mail protocol and networking; information about Exchange Server® and events.

[37] The SME should be ready for heavy involvement when needed, not merely someone whom the analyst is referred to, and whose official responsibility does not include support of the SNA. Even officially supporting the project on paper is not sufficient if such support always takes a back seat to operational priorities, since there always seems to be many more of those priorities than there is time for.

- On the other hand, email administration, Exchange Server®, and directory services have large amounts of technical knowledge behind them, which can also impede the search for information.

  Because of the sometimes voluminous documentation, the prospects of identifying the specific information to resolve a particular issue could be quite small without expert guidance. For those not in the field, it is sometimes unclear that the right information is being looked at, even after it has been located.

- Access to subject matter expertise is required in all aspects of implementing data gathering and preparation on the mail servers, the options for which are speculated in Section 5.1.
  - Expertise is needed to confirm whether suitable capability actually exists on the servers, as well as the expertise and effort needed to utilize it for generation of SNA input data.
  - Since implementing solutions on the servers can be a sensitive issue, subject matter expertise would also inform an estimate of the degree of imposition at the servers to do this, which affects the likelihood of approval.
  - If in-house implementation of such a solution is required, the need for Exchange Server® expertise will also be likely.

- It is conceivable that cost and approval barriers rule out all options to data preparation except for the culling of data from the tracking logs only, with possible utilization of downloaded GAL data to identify users rather than online access to DWAN's AD. An informed assessment of the approach is then needed, based on expert knowledge of the following:
  - The circumstances under which the different combinations of identity information arise in the tracking log files
  - Why the X.500 data is missing disambiguating middle initials, while they seem to be present in at least some of the X.400 data.
  - The limits on reliability of any empirically based identification scheme. Even if great effort was expended to concoct byzantine schemes that border on forensics or reverse engineering, and that seem to resolve all the relevant identities in the example log, there is no guarantee that they will work for similar data in the deluge of logs for the actual SNA. Nor is there guarantee that new forms of cryptic data will not be encountered. This is particularly true in view of the fact that log data from OGDs have not been studied. The unreliability of empirically crafted identification schemes is the consequence of the bottom-up characterization, since it is not necessarily known why the data is generated in the manner that it is. Though Microsoft™ indicated that not many people today know the inner workings of the older server logs, a SME would have more experience on which to base opinion about the reliability of an empirical identification scheme.
  - As a basis for developing an empirical identification scheme, it is advisable to get a more experienced opinion about how representative are the example log files (or to use actual log files from the servers of interest, if they are available at

the time the scheme is being devised). For Polar Guardian's SNA, the sample log file for Exchange Server® 2003 was especially in need of this appraisal.

- As per the above comment on expert access to arcane information, an email administrator would likely have access to crucial documentation, such as the event definitions for Exchange Server® 2003. This is necessary to select the proper records from the log file. It also provides clues about the problems that need advance recognition; for example, the questionable representativeness of the 2003 sample log should be recognized *before* any effort is spent on an empirical identification schemes. The event definitions only became publicly accessible during the final drafts of this technical note; the analysis for which it is needed should not be so dependent on such happenstance.

5.5 Subject Matter Expertise for Options Assessment

- Vetting of commercial tools would have been far more efficient and lucid with access to the proper knowledge about email and directory services.

  Vendors of tools such as those vetted are usually speaking to SMEs. Not having that background is a communication barrier.

- For commercial solutions that qualify as candidates, subject matter expertise may be needed to determine whether imperfect robustness (Section 5.3) points to a fundamental limit in identification.

  For the relevant identities (those in records for non-intermediate hops) that do not appear to represent users at all, tool performance will indicate whether they actually do not lend themselves to any kind of identification method. A SME might be able to determine whether this is a fundamental limitation in that there simply isn't enough data to resolve identity even with the use of the AD, regardless of the tool or method. If so, it is an upper bound on the accuracy of the data for the SNA.

  Such a determination has ramifications for the home-grown options also. No further effort should be spent on the impossible identifications. This should simplify home-grown solutions, decrease development time, and dispel their disadvantage relative to commercial solutions.

- On the other hand, subject matter expertise may also inform the determination that such an upper bound on accuracy is not as solid as it seems e.g. because bad data can be circumvented by tracking messages across server logs and combining the records.

  The apparent impossibility of identification may actually be relative, and may depend on the cost and amount of effort that can be afforded on forensics. For example, the tracking logs may contain enough data to mine and reconstruct message pathways from server to server, thus allowing sender/recipient to be identified from some records even when they are missing from others. This generalizes the scheme of matching incoming and outgoing messages (Section 4.5.3). A quick search shows that such tracking is almost certainly available on Exchange Server® software. (With this functionality in mind, and the knowledge of multiple hops per email delivered, the corroborating descriptions in the

tracking log specifications of Annex F become evident). The SME would have a much better idea of:

- The robustness of using tracking to completely identify sender and recipient
- How easy it is to get data in the form required for the SNA from the results of such tracking by the server.
- The tractability of tracking a large volume of email in the context of DND's large email system
- In consultation with vendors e.g. in Section 4.2 and Annex D, how likely it is that a robust commercial tool relies on tracking, either built-in or from the server. Such reliance exposes them to any limitations in message tracking that may apply within the DND email environment.

The practicality of developing such tracking functionality in-house in a reasonable amount of time is questionable because it would require intimate familiarity with the Exchange servers, tracking logs, and algorithms and data structures for high-speed matching over a large volume of data. For commercial tools, however, it isn't clear what is impractical once data is filtered, pre-processed, and read into a relational database, such as that required by the tools vetted.

A potential source of intractability in DND is the need for the logs of intermediate servers in the reconstructed message pathways, especially if such pathways are not deterministic (an SME would likely know if they are). In the worst case, the paths are completely unpredictable and the logs for all the servers would be needed. This implies a large number of servers to monitor, even if there are a small number of interesting users and a small number of servers hosting them. Based on subject matter expertise, the tractability of the required amount of data and processing needs to be checked with the scope of the project.

- Depending on the technical details behind the various user identification problems and the various approaches to preprocessing, subject matter expertise may be needed to assess the impact on the resulting SNA data.

  A generalization of the previous point is that there may be identities that require varying degrees of sophistication and resources to resolve correctly, which means that the various tools and approaches will have different degrees of identification robustness and feasibility. This applies to solutions that are home-grown, use commercial tools, or a combination of the two. For example, solutions that simply discard records containing hard identification problems may be quite feasible[38], but not robust against many of the vagaries of the identification data. Expertise in Exchange Server® operation could inform the assessment of how severe are the effects of lost data by considering the *circumstances* that give rise to the records that are not handled (or worse, improperly handled) by said solutions rather than considering just the quantity of erroneously processed records. For example, 10% loss in data may be acceptable if it was randomly distributed, but can change the patterns in the SNA graph if the cause is systematic. If the erroneous records actually misidentify sender

[38] In terms of cost, time, and technical difficulty.

or recipient, a SME might also be able to estimate the seriousness of the SNA data pollution by considering the technical details of how the misidentification is made[39].

5.6 Feasibility of Email SNA

The lessons learned has so far dealt with the challenges in planning SNA data collection and preparation, and in the next level of detail, the challenges and unknowns in the candidate approaches to their solution. At this point, an SNA within DND, let alone externally, sounds very difficult.

Many of these challenges, however, stem from the particular circumstances of the Polar Guardian SNA. If this had to be boiled down to one all-encompassing point, it would be the fact DIMEI was not approached for *active* involvement with the planning at the outset. This oversight is not difficult to understand. Without awareness of how involved it can be, why *wouldn't* one initially forge ahead, gather data, preprocess it, and analyze the result? This led to the discovery of the bottom-up problem of trying to generate meaningful SNA input from possibly not the best source, with a just an articulation of how it could be done otherwise, along with some risk identification.

Getting buy-in from a level high enough in DIMEI to commit the necessary SME resources to SNA could completely change the outlook. If options for data gathering at the servers were well understood, it is possible that all the challenges and questions related to user identification in the tracking logs could simply become irrelevant. In fact, it is difficult to see how a SME would *not* be able to advise on the most suitable method of data collection and help with the permissions in setting it up. The inclusion of technical subject matter expert in the SNA effort would avoid the situation where the SMEs respond primarily to operational priorities, with responses to queries from peripheral research efforts by unknown analysts on an opportunistic basis. Operational priorities often mean that such sporadic responses can do little more than factually answer the technical questions as-posed, sometimes after several iterations of clarification, rather than providing contextual information and advice on the approach (for example, on alternatives to generating information from the tracking logs).

In this study, DIMEI was found to be a key player, technically and otherwise. The organizational and interpersonal dimensions of engaging key players are not technical, but are clearly recognized in an explicit manner in [10].

For an interagency SNA, the greater technical diversity, and organizational and political complexity, mean that having buy-in from a high enough level is even more important to ensure that the relevant information stewards and technical SMEs support the effort.

5.7 Considerations for an Interagency SNA based on Email

The prospects of incorporating OGDs into an email-based SNA can be expected to depend on the data collection method. If it deploys applications on their mail servers, OGDs may be quite sensitive about it. If it merely requires access to their mail server logs and access to their AD for

[39] Indeed, the feasibility and confidence of such an estimate would be in the province of the SME.

user identification, the prospects are likely better[40]. Liaising with the OGDs is needed, however, to get a more concrete idea of the likelihood for approval. Subject matter expertise is needed to determine if these two scenarios are indeed the only main options.

Subject matter expertise is also needed to anticipate the severity of the following challenges, and the expertise and tasking to resolve them.

- The challenges in integrating the data across separate domains need to be determined
- The challenges arising from possibly dissimilar mail systems, networks, and configurations thereof, need to be anticipated.
- Liaising with the OGDs from a position of technical expertise is needed to assess the extent of such dissimilarities. Whether disparateness of systems/networks within the same OGD poses additional challenges also needs to be determined.

If it turns out that the challenges to such an SNA are beyond the scope of a given project, a completely survey-based SNA may be preferable.

5.8 Survey-Based SNA: Motivations

Though an SNA based on electronic communication can be appealing for its objectiveness, a survey-based SNA provides flexibility in the design of the questions to elicit data about different kinds of relationships e.g. the different circumstances that prompt different kinds of communications, and the kinds of information conveyed. While most of the time in Polar Guardian's SNA was focused on getting the right input from logs on electronic communication, two suggestions were kept in mind for the follow-up survey phase discussed in Section 1.2.

- If possible, engage participants with SNA graphs and analysis results from an email-based SNA prior to conducting the survey.
- If it is appropriate for the culture of the organization(s), offer reasonably enticing incentives for returned surveys. It has been found that gifts work well for this. There is no second chance, and missing information affects the patterns in the social network, and potentially the conclusions that arise from its analysis.

[40] From the perspective of policies/procedures pertaining to security and email administration. As mentioned for intra-DND email, the ethics and legality also need investigation.

6 Conclusions and Recommendations

Though the email-based SNA in this project was not carried out to completion, the preliminary investigations under Polar Guardian can be considered a success in establishing a baseline knowledge about what is involved in conducting such a study. Going into this effort, there was very little information about what could be expected in obtaining and preparing data for such an SNA outside of a completely controlled common operating environment for an experiment. Technical and administrative/policy requirements and challenges were identified, as were possible approaches to their solution. Possible trade-offs of the approaches were identified with respect to uncertainties in technical feasibility, acquisition of assets, installation approvals, and delay. Furthermore, areas were identified in which expert consultation and investigation are needed, especially concerning solutions implemented onto the email servers, and challenges/unknowns in compiling data on interagency emails.

The overarching lesson that can be taken away from the Polar Guardian effort is the need for the active involvement of an SME on email administration, and possibly on directory services for local area networks. It is also important that the SME have visibility into the DND email system e.g. the SME would have had involvement with DIMEI, or ideally, would come from DIMEI. Opportunistic discussions with IT staff that have visibility into the areas of email administration and directory services indicate that these knowledge areas are vast. Obviously, not all of that knowledge is needed to obtain SNA data; it isn't clear, however, how much of that knowledge is required. Despite the technical challenges and unknowns reported herein, the bridge of technical knowledge and interdepartmental cooperation that would be enabled by SME involvement could invert the outlook on the SNA's feasibility from one containing many challenges, unknowns, and possible outcomes to one containing a few (or a single) simple way(s) forward, with minimal uncertainty as to the resources, effort, and time required, and a clear understanding of the caveats associated with the resulting data.

In an email-based SNA, there is a serious need for such technical expertise for all aspects of data preparation, including the assessment and implementation of the three broad options: (1) direct use of partial information in the tracking logs, with possible assistance from directory services, and the impact of irresolvable identities on the SNA; (2) commercial tools that claim to be able to generate the required data, either from the servers or from tracking logs, with possible assistance from directory services; and (3) the ability to generate the required data using email server capabilities.

Within DND, the need for this expertise will be in the area of Microsoft Exchange Server®. With respect to option (1), such expertise is needed to assess and/or mature/complete the solutions and trade-offs described in this study. With respect to options (2) and (3), such expertise is needed to fully understand the need for, and implications of, deploying various applications onto DWAN machines and/or email servers, and/or access to directory services. This expert understanding is essential to establishing a case for their purchase and/or deployment, if necessary.

Technical expertise in email administration is also needed to scope out and resolve the challenges in obtaining and using data from multiple agencies. Depending on discussion with email administrators in the other agencies, and the expertise that can be leveraged from them, it is

possible that such expertise may have to extend beyond Microsoft Exchange Server® to other email servers, and to IT standards pertaining to email in general.

The involvement of the SME cannot be superficial. He/she should be readily accessible, as needed, and take an active role in the solution to the data acquisition and preparation. It is recommended that the SME be a member of the team conducting the SNA, at least for the initial stages for data acquisition, and preferably beyond.

With limitations such as those encountered in Polar Guardian, however, the time frame for the SNA needs to be expanded to properly assess the suitability of commercial solutions e.g. through demonstrations that target to the project's needs, or (especially) to develop solutions to the user identification problems in home-grown solutions.

With respect to commercial solutions, an assessment is also needed of CFEC's ability to meet the prerequisites of those applications in a timely manner, such as availability of database software, and ready access to directory services such as Active Directory. Delays in approval, purchase, and/or installation need to be taken into consideration.

With respect to home-grown solutions, it may be the case that within the project limits for time or resources, not enough insight can be gained into causes of, and solutions to, the user identification problems, or that the solutions are not feasible. The social network analyst may simply have to accept that a portion of the communications will not be reflected in the SNA. SME input could shed light on how adversely the SNA would be impacted by this, depending on how systematic is the lost data. The analysts will then have to decide how important are such inaccuracies, depending on the purpose of the SNA and the circumstances that lead to lost data.

With the current lack of subject matter expertise, solutions that need to be deployed on the email servers are not likely to be technically feasible in the short term. The administrative delays in their approval, and the risk of denial, are also expected to be significant.

The current outlook on the three factors above (time expenditures to assess candidate products, potential inaccuracies in the email traffic volumes from home-grown solutions, and challenges to server-hosted solutions) may change significantly with new knowledge, such as input from an SME. It is possible that the aforementioned expansion of the time frame in particular could be made more concrete with expert technical input, and/or that it need not be significantly more than initially conceived if an SME is involved with the implementation or deployment.

A significant lead-time should be scheduled for: (1) obtaining approval of the legality and ethicality of accessing the data on email communications; (2) obtaining permission for, and arranging, access to the data; and (3) identifying the users to be studied. Within DND, all of these seem to be plausible, if not quick. For interagency communication, the feasibility and time frames for these requirements need to be assessed in consultation with the relevant authorities in each organization. It is possible that authorities outside of the participating organizations may also need to be consulted regarding (1). With respect to (1) and (2) (as opposed to the previously mentioned technical approvals for deployment of applications), subject matter expertise in email administration and/or directory services is expected to be important in establishing the case for their approval.

The discussion thus far has focused on identifying and addressing the front-end challenges to corporate SNA based on email, and to a limited degree, interagency SNA based on email. It would be fitting to also discuss what SNA can do for DND/CF, and hence why such additional lengths are worthwhile[41]. Section 1.1 describes the extreme situation in which it is not known how closely actual communications follow formal protocols, in the context of interagency SA. Within the military, however, there can be a wide variation in the level of detail to which formal protocols are specified, beyond SA to command and control, and depending such factors as a particular commander's personality [10]. SNA is a tool to reveal actual communications patterns in order to sanity check and streamline operations. Analysts are not left relying solely on patterns suggested by protocol, which themselves might be rather open-ended, and the adherence to which can only be assumed[42]. Such ground truth is important for situations requiring timely shared SA, completeness of information, and timely response.

In the domain of joint, multinational and interagency experiments and exercises[43], SNA has already been used to inform concept development[44], albeit in a controlled experimental communications environment. Extending this into the operational domain within DND and beyond would bring insight not only to higher fidelity exercises that are conducted on operational equipment/facilities, but also to operations itself. An SNA could suggest where to investigate further in troubleshooting unexpectedly untimely response. Operations can also be compared with exercises to improve the fidelity of the latter. A final example of SNA's utility is to inform the development of protocols for a liaison officer in a operations centre to reach back to his/her home organization, based on the information flows therein.

Both exercises and operations can be studied to validate the thinking behind concepts of operation, not only by characterizing specific communications pathways of initial interest, but also by revealing unexpected features. One example is to identify highly connected nodes that may be critical failure points, and that could benefit from planned redundancy e.g. in staffing. Another example comes from a recent study revealing no communication flows between Marine Security Operations Centre East and Government of Canada Operations Centre, thus indicating a lack of formal mechanisms for their interaction [11][45].

[41] These considerations also apply to the more complete SNA based on electronic communications, including phone communication, as mentioned in Section 1.2.
[42] In fact, doctrine in particular is only meant to represent the guiding default conduct for a situation, to be overridden whenever an alternative is deemed more appropriate.
[43] Becoming increasingly important in the current day push for a comprehensive approach to CF operations [12].
[44] For MNE 4, the SNA included all manner of electronic interaction [4][5]. SNA was also intended for MNE 5, but was not completed due to complications in maintaining the resourcing of the SNA with key personnel. Chat data from MNE 4 was also analyzed outside of the MNE 4 purview [13], in conjunction with email data from an interagency Command Post Experiment known as Pegasus Guardian [14][15][16].
[45] The communication picture in the latter example is more of a cross between protocol and reality, since the data was gathered from the operators and documents rather than based on actual communications.

References

[1] Huebert, R. (2005), Renaissance in Canadian Arctic Security?, In D.L. Bashow (Ed.), *Canadian Military Journal*, 6(4), 17--29.
Available: http://www.journal.forces.gc.ca/engraph/Vol6/no4/PDF/04-North1_e.pdf
(Access date: 22 Feb. 2008).

[2] Arctic Surveillance Interdepartmental Working Group (ASIWG) (2005), *Arctic Issues Information Collection Project: ASIWG 23-24 November 05 – Update* (Yellowknife).

[3] Cross, R. and Parker, A. (2004), The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations. Boston, MA: Harvard Business School Press.

[4] USJFCOM Joint Experimentation Directorate (J9) (2006), Multinational Experiment 4 – Final Report. US Joint Forces Command, prepared with contributions from Australia, Canada, Finland, France, Germany, Sweden, United Kingdom, and NATO.

[5] Farrell, P.S.E et al. (2006), Multi-National Experiment 4 on Effects Based Approach to Operations: CFEC Analysis Report (DRDC Ottawa TR 2006-230) Defence R&D Canada – Ottawa.

[6] Resnick, P. (2001), RFC 2822: Internet Message Format (online), Network Working Group, http://www.rfc.net/rfc2822.html (Access date: 26 Oct. 2006).

[7] de Nooy, W., Mrvar, A., and Batagelj, V. (2005), Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences series). Cambridge, New York, Melbourne: Cambridge University Press.

[8] (2008), PivotTable Reports 101 (online), Microsoft™, http://office.microsoft.com/en-us/excel/HA010346321033.aspx (Access date: 22 Feb. 2008).

[9] Hill, S. and Woodill, G.E. (2007), DMP Strat/CORA Research Ethics Board: Two Year Performance Evaluation, (DRDC CORA TN 2006-12) Defence R&D Canada – CORA.

[10] (Revised 2002), NATO Code of Best Practice for C^2 Assessment. DoD Command and Control Research Program.

[11] Carson, N. and Smith P. (2006), Marine Security Operations Centre (MSOC) East . Work and Information Flow Analysis: Phase 1 Report (U), (DRDC CORA TR 2006-11) Defence R&D Canada – CORA.

[12] Leslie, A., Gizewski, P., and Rostek, M. (2008), Developing a Comprehensive Approach to Canadian Forces Operations, Canadian Military Journal, 9 (1), 11--20. Soft copy accessible at http://www.journal.forces.gc.ca/vo9/no1/04-leslie-eng.asp (Access date: 24 July 2009).

[13] Lisa Rehak et al. (being finalized as of July 2008), Final Synthesis Report for Social Network Analysis of Joint Interagency Multinational & Public (JIMP) Operations (DRDC Toronto CR 2008-XX), Humansystems Incorporated, Guelph, ON

[14] Canadian Forces Experimentation Centre (2007), Preliminary Quick Look Report - Command Post Experiment Pegasus Guardian.

[15] Allen, D (2009), Comparison of Planned and Observed Processes for Major Event Security - Results from Pegasus Guardian 1 (DRDC CORA TN 2009-024), Defence R&D Canada – CORA.

[16] Allen, D., Chow, R., Trinh, K., and Farrell, P.S.E. (awaiting publication as of July 2009), Pegasus Guardian 1 Experiment - Analysis Report (DRDC CORA TR 2009-XX) , Defence R&D Canada – CORA.

[17] (2008), Exchange Server 5.5 Product Documentation: Maintenance and Troubleshooting (online), Microsoft™, http://www.microsoft.com/technet/prodtechnol/exchange/55/proddocs/server4.mspx (Access date: 22 Feb. 2008).

[18] (2008), Message Tracking Logs field descriptions in Exchange 2000 Server (online), Microsoft™, http://support.microsoft.com/kb/246965 (Access date: 22 Feb. 2008).

[19] (2008), Message tracking event IDs in Exchange Server 2003 (online), Microsoft™, http://support.microsoft.com/kb/821905, 27 Oct 2006 (Access date: 22 Feb. 2008).

Annex A Overview of Pajek

The aim of this annex is to provide a quick overview of Pajek, a freeware application that performs social network analysis. In particular, the various commands of Pajek are only partially described and the discussion focuses on the utility of Pajek. The readers interested in learning how to use the program are referred to the user manual which is available online at the following URL: <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>.

Pajek is a program for the analysis and visualization of large networks. This program was developed by Vladimir Batagelj and Andrej Mrvar of the University of Ljubljana in Slovenia. Pajek evolved largely since its first version came out in November 1996. The program is a freeware and its latest version (Pajek 1.14) can be downloaded at the following URL: <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>. Although Pajek can be used for representing chemical molecules, its main utility is for social network analysis, the goal of which is to identify and interpret pattern of social ties among actors.

From a SNA perspective, a network is equivalent to a mathematical graph - composed of vertices and links between some of these vertices - with additional information associated to the vertices and links (e.g., vertices label and type of link). Pajek deals with various objects that can be associated with a network: the network itself, partitions of vertices, permutation of vertices, clusters of vertices, hierarchies of vertices, and numerical vectors. These objects are defined as follows:

- Network: Set of vertices and links between vertices with possibly additional information associated with the vertices and links.
- Partition: Object that associates each vertex with a given class of vertices. Partitions may specify structural properties or attributes of vertices.
- Permutation: A set of rules modifying the ranking of an ordered set of vertices.
- Cluster: Subset of vertices. A given class of vertices constitute a cluster but the reverse is not true. In other words, the vertices belonging to a same cluster do not necessarily belong to the same class.
- Hierarchy: Ordered subsets of vertices.
- Numerical vector: Use to associate a set of numerical properties to each vertex.

Pajek offers the possibility to create, modify, transform, and visualize all these type of objects. It is also possible to write a program that would display the dynamic of the visualized network. Finally, Pajek can be used to obtain various properties and statistics on the created objects.

Figure A-1 displays Pajek's interface. On the window appears drop down menus to access all loaded objects: networks, partitions, permutations, clusters, hierarchies, and vectors. All the loaded objects can be transformed and analyzed using the menu available at the top of the window.

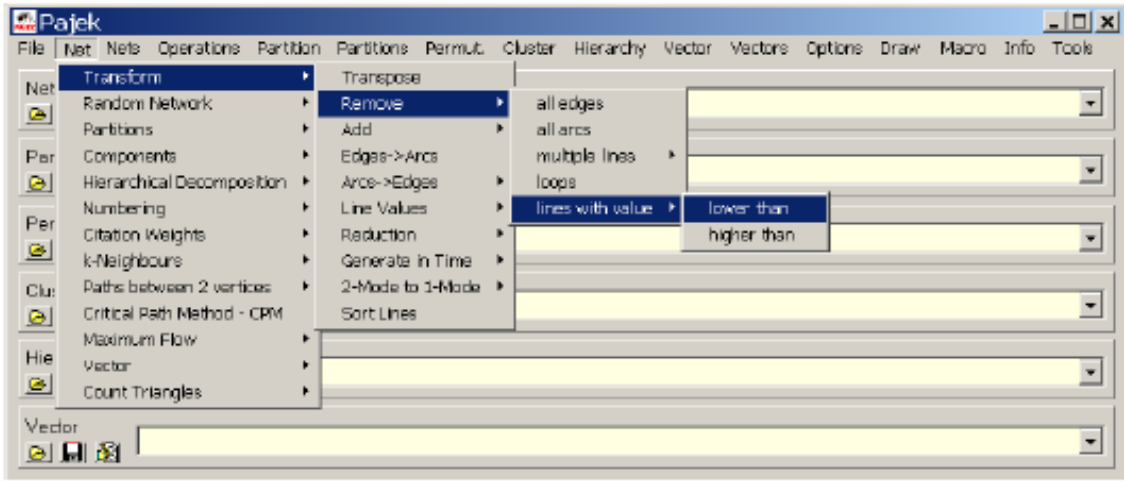

Figure A-1. Pajek's main interface window.

Although networks can be created manually within Pajek, it is also possible to create a network from a DOS text or ASCII file. A generic input file will have the following structure:

```
*Vertices 5
        1 "Josh"      0.15    0.25    0.5
        2 "Phil"              0.35    0.7     0.5
        3 "Bill"              0.55    0.11    0.5
        4 "Karl"      0.75    0.65    0.5
        5 "Anna"      0.9     0.45    0.5
*Arcs
        1       5       0.8
        2       5       0.7
        3       5       0.3
        4       3       0.6
        5       4       0.8
*Edges
```

Two different types of links can be entered in Pajek: arcs or edges. Arcs are directed links going from one vertex toward another one. Edges are undirected links. In the example above, the network is composed of 5 vertices. A label is assigned with each vertex (e.g., "Josh"). Three numbers between 0 and 1 are also used to specify the 3D-coordinate of each vertex. Below the vertices data are the arcs data. Each arc is described by first specifying its starting vertex, then its ending vertex, and finally an attributed weight. Similar data would be used to specify the edges. Note that the vertex label, vertex coordinates and the weight of the links are all optional data. Pajek possesses algorithms to select optimal locations for the nodes. In particular, the user can select the coordinates of the vertices in a way to minimize the number of links crossing each other or to impose a minimum distance between unlinked vertices (it is largely accepted that the visual distance between vertices should be inverse to the number of links between them, strongly linked vertices being near each other).

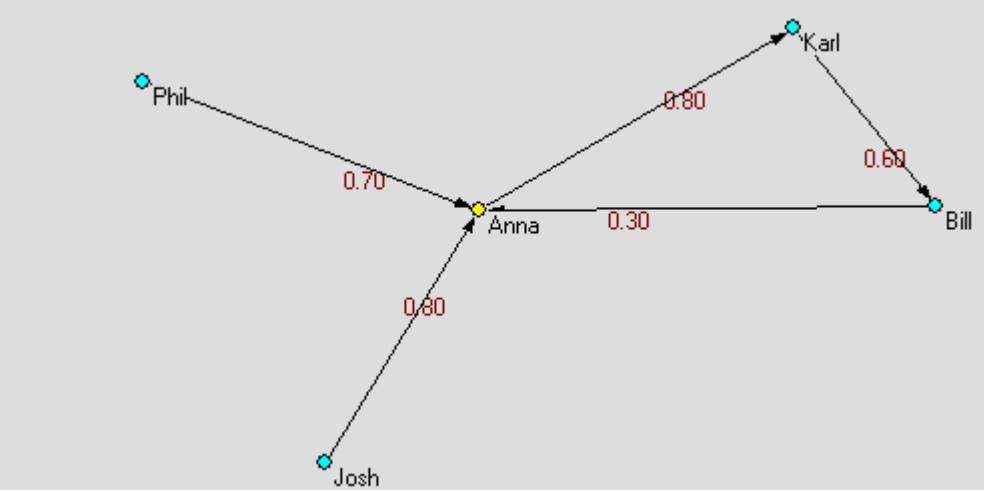

Figure A-2. Network as displayed by Pajek.

Figure A-2 shows the graph as displayed by Pajek after optimizing the node location. The weight associated with each link is displayed in red. Note that the nodes are also partitioned according to gender. The partition is displayed by the color associated with the node: yellow for female, blue for male.

In addition to displaying the partition using color code, Pajek can also display numerical properties associated with the vertices by varying the size of their associated node. Figure A-3 displays a network where the following values were associated with each node: 'Josh' has a value of 10, 'Phil' a value of 15, 'Bill' a value of 20, 'Karl' a value of 12, and 'Anna' a value of 20.

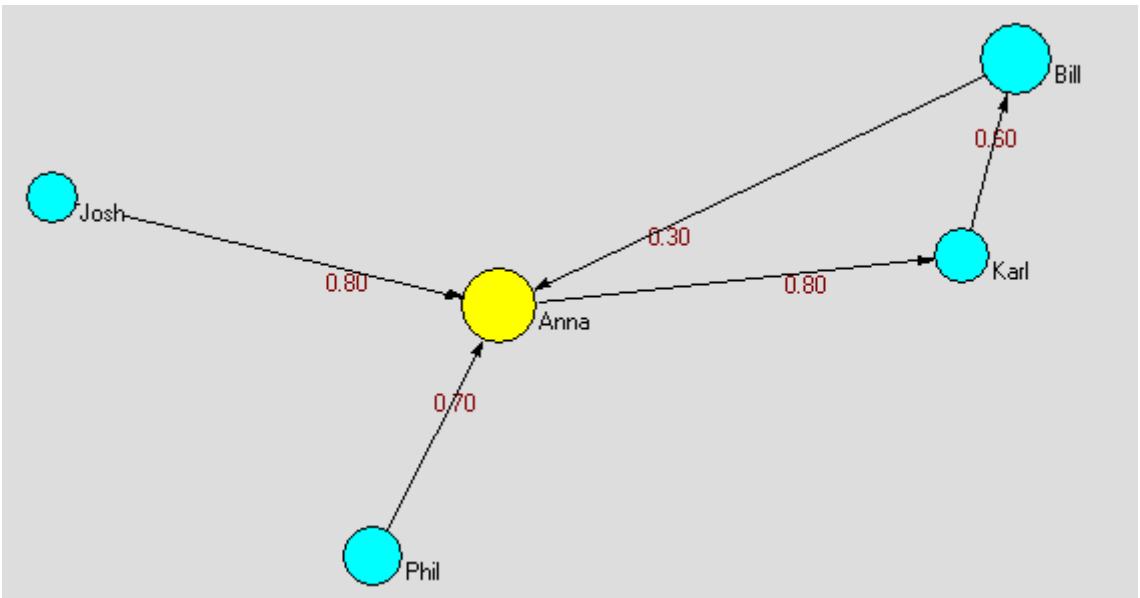

Figure A-3. Network with partition and vector as displayed by Pajek.

Pajek also offers tools for the visualization of large networks. It is possible to merge all the vertices pertaining to the same partition into a single vertex. This way, a more global view of the network is obtained; only the connections between the different partitions is displayed (this operation is called a *global view*). Another possibility is to display only a subset of the network, deleting all other vertices and links outside of this sub-network (this operation is called a *local view*). It is also possible to merge these two operations: all the vertices, except those pertaining to a given sub-network, are merged into aggregate vertices based on the partition to which they pertain (this operation is called a *contextual view*).

In addition to visualization tools, Pajek offers various tools for analyzing networks. In particular, various measures of centrality of the vertices can be computed (refer to the resources in the Bibliography for a discussion of centrality measures). The overall structure of the network can also be analyzed. For example, Pajek can find important links, the removal of which would break the network into unconnected parts. Pajek can also be used to investigate the strong components of a network. In the example above, Anna, Karl, and Bill form a strong component because it is possible from any vertex of this sub-group to move to any other vertices of this sub-group while following the directed arcs. Finally, Pajek can be used as a simulation tool. For instance, the spread of a disease or of information among a group of people represented by vertices can be simulated.

Annex B Formulae for Soft Limited Weighting of Multi-recipient Emails

One need only consider the less relevant broadcast emails that one receives to realize that a single email sent simultaneously to N=20 recipients is unlikely to be worth as much as 20 different emails sent individually to different people, at least in terms of meaningful communication that is indicative of a close relationships. It is necessary to count the 20-recipient email as being worth somewhat less than 20 individual emails, so the contribution to the SNA connection to each recipient will be less than a single email i.e. a fraction of one email. A means is needed to quickly assign such a attenuating scale factor, or weighting, to the overall number of 20 recipients to arrive at the equivalent number of single-recipient emails. This will commensurately attenuate the contribution to the SNA connection to each recipient.

This annex focuses on the algebraic form of formulae for soft limited weighting. It does not cover further elaboration on the motivation for such weighting.

Different weighting examples can be devised if one regards emails with recipient lists longer than some threshold (say N_0=30 people) as contributing no additional information about relationships between individuals. Under such a weighting, the value N_{Tot} of an N-recipient email approaches the equivalent of N_0 single-recipient emails as N increases, but never reaches N_0. For example, the total weight $N_{\text{Tot}}(N)$ for all N recipients can defined as the lesser of $N_{\text{Tot}}(N)=N$ and $N_{\text{Tot}}(N)=N_0$, with a soft transition where the two cross over. Formulas can be readily borrowed from semiconductor physics e.g., $N_{\text{Tot}}(N)=N_0N/(N_0+N)$ [46], which soft-limits at N_0=30 (Figure B-1). This implies a per-recipient weight of $N_{\text{Tot}}(N)/N=N_0/(N_0+N)$.

A more flexible soft limiting can be devised based on logarithmic Bode plots for electronic circuits, where frequency response follow straight lines, with soft transitions as they cross over. To soft limit at N_0, it can be shown that $N_{\text{Tot}}(N) = N_0 - \log_b(1 + b^{N_0 - N})$, where b>1 results in soft limiting for values close to 1, and sharp limiting for large values e.g. 20 (Bode plots use 10 by default, but the curves tend to have sharp knees). Figure B-2 shows a variation of this (explained below) for b≅1.2. The analyst can experiment with b until he/she obtains a curve of diminishing returns that is felt to be appropriate.

Figure B-2 incorporates a correction factor to eliminate a slight nonzero offset in N_{Tot} at N=0. The offset results from the fact that $N_{\text{Tot}}(N)$ only approaches the bounding lines $N_{\text{Tot}}=N$ and N_{Tot}=30 *asymptotically*. In Bode plots, the offset is rendered insignificant by the large b value, but a correction factor should be considered here because b is intended to be adjustable. The correction consists of vertically compressing the curve toward the bounding line N_{Tot}=30 by an amount needed to have $N_{\text{Tot}}(N)$ cross the origin. This amount is determined by evaluating the vertical offset at $N_{\text{Tot}}(N)$ at N=0. In formal terms, the compression consists of a few linear geometric transformations: A downward translation so that N_{Tot}=30 aligns with the x-axis, a vertical compression by the required amount, and then an upward translation to reverse the

[46] For sake of attribution, this formula is from electron mobility and velocity saturation. Awareness of such origins is not necessary to appreciate the graphical behaviour.

downward translation. The result is $N_{\mathrm{Tot}}(N) = N_0[1 - \log_{(1+b^{N_0})}(1 + b^{N_0 - N})]$, shown in Figure B-2. For simplicity, the following discussion uses the simpler form $N_{\mathrm{Tot}}(N) = N_0 - \log_b(1 + b^{N_0 - N})$.

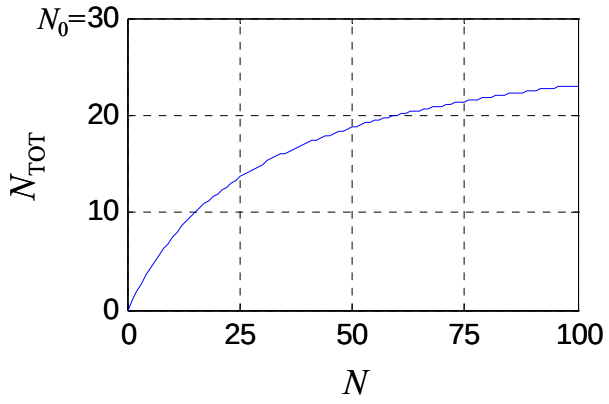

Figure B-1. Example of sublinear[47] function N_{Tot} as a function of N which soft-limits at N_0=30: $N_{\mathrm{Tot}}(N)=N_0N/(N_0+N)$

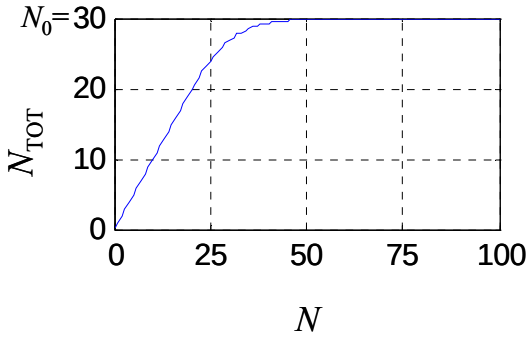

Figure B-2. Example sublinear function $N_{\mathrm{Tot}}(N) = N_0[1 - \log_{(1+1.2^{N_0})}(1 + 1.2^{N_0 - N})]$, which soft-limits at N_0=30.

[47] There are rigorous ways to define sublinearity, but here it refers to the behaviour of a single-input, single-output function in the upper-right quadrant of a Cartesian graph (the only region of interest). The goal of the function is to represent diminishing returns, so the slope is always positive, and always decreases as the independent variable increases.

The algebraic generation of a curve of diminishing returns that asymptotically approaches a limit can be generalized very intuitively, in case the analyst wants greater flexibility in the sharpness of the knee. As mentioned, such a curve asymptotically approaches the bounding lines $N_{Tot}=N$ for $N<N_0$, and $N_{Tot}=N_0$ for $N>N_0$. The following discussion assumes that the regime of interest is the upper-right quadrant of a Cartesian graph (independent and dependent variables are positive).

A function $f(N)$ can be devised to approach bounding curves $g_1(N)$ and $g_2(N)$ by composing it as $f(N)=amp^{-1}\{amp[g_1(N)]+amp[g_2(N)]\}$, where *amp* is a superlinear[48] "amplification" function. In regimes of N away from their cross-over point, $g_1(N)$ and $g_2(N)$ are expected to differ nontrivially, and *amp*'s superlinearity ensures that $amp[g_1(N)]+amp[g_2(N)]$ is dominated by one of the two terms. The subsequent application of amp^{-1} then yields an $f(N)$ that approximates the larger of $g_1(N)$ and $g_2(N)$. Since $amp[g_1(N)]+amp[g_2(N)]$ are additive, however, $f(N)$ will always be *above* both $g_1(N)$ and $g_2(N)$. It is straightforward to show that $N_{Tot}(N)=g_1(N)+g_2(N)- f(N)$ will be *below* both $g_1(N)$ and $g_2(N)$, and asymptotically approach the lower of the two, which is the desired behaviour.

For the Bode-based curve of Figure B-2, $N_{Tot}(N)$ can be written $N_0 + N - \log_b(b^{N_0} + b^N)$, which shows that $g_1(N)=N$, $g_2(N)=N_0=30$, $amp(N)\equiv b^N$, and $amp^{-1}(N)\equiv log_b N$. Figure B-3 shows another example with $amp(N)\equiv N^3$. Figure B-1 is of a different algebraic form, but is well approximated by $amp(N)\equiv N^{1.7}$; similar numerical dynamics prevail in that the limiting boundaries are transformed into another domain (via arithmetic reciprocation, as can be easily checked), in which one term dominates in an addition, before being transformed back to an approximation of the dominant boundary.

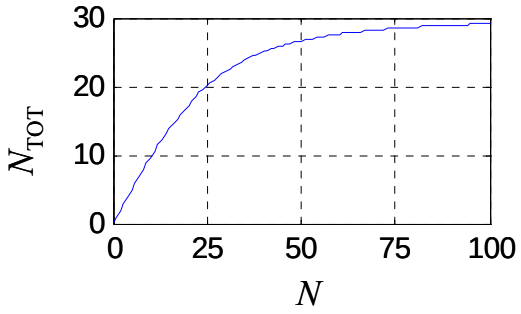

Figure B-3. Example sublinear function $N_{Tot}(N) = N_0 + N - [N_0^3 + N^3]^{1/3}$, *which soft-limits at* N_0=30.

[48] Here, *superlinear* refers to the behaviour of a single-input, single-output function in the upper-right quadrant of the Cartesian graph. The slope is always positive, and always increases as the independent variable increases.

Choosing a power function $amp(N)=N^M$ (Figure B-3) is algebraically simpler than the exponential function of Bode-based curves because $amp(0)=0$, thus ensuring that there is no offset at $N=0$, and no correction factor is needed.

While there is some judgement on how to shape the curve with which to weight a multi-recipient email, the weighting should be consistent with a single-recipient email. The above monotonic total weightings closely approach 1 as $N \rightarrow 1$, and 0 as $N \rightarrow 0$.

For an N-recipient email, once the curve for the total weighting N_{Tot} is determined, the per-recipient contribution to the SNA connection is simply N_{Tot}/N (Figure B-4).

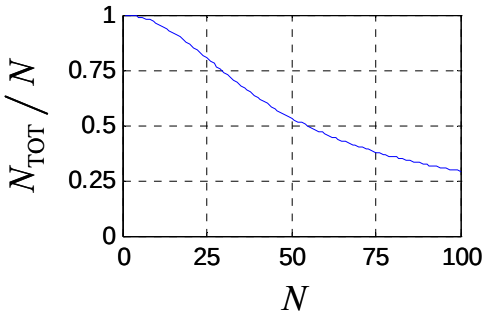

Figure B-4. Per-recipient contribution to SNA connections for the N-recipient email of Figure B-3.

Annex C Interagency Stake Holders

The following stake holders were culled from the minutes for ASIWG 16-17 May 2006.

C.1 Federal

Canada Border Services Agency

Canadian Coast Guard

Canadian Ice Service

Canadian Security Intelligence Service

Canadian Space Agency

Citizenship and Immigration Canada

Department of National Defence:
CFEC, Canada Command, Strategic Joint Staff, DRDC, JTFN

Department of Fisheries and Oceans

Environment Canada

Foreign Affairs and International Trade

Health Canada

Indian and Northern Affairs Canada
Northwest Territories (NT), Nunavut (NU), Yukon (YT)

Industry Canada

International Polar Year Federal Program Office

Justice Canada

Northwest Territories Federal Council

National Energy Board

Natural Resources Canada

Public Safety and Preparedness Canada

Parks Canada

Public Health Agency of Canada

Royal Canadian Mounted Police:
Northwest Region Immigration & Passport Section
Divisions: G (NT), M (YT), V (NU)

Service Canada

Transport Canada

Yukon Federal Council

C.2 Provincial

Government of Northwest Territories:
Emergency Measures Organization
Intergovernmental Relations
Justice

Government of Nunavut:
EMO (Nunavut Emergency Management)
Intergovernmental Affairs

Government of Yukon:
Emergency Measures Organization
Intergovernmental Relations
Justice

C.3 Ethnic

Nunavut Tunngavik Inc.

C.4 Municipal

Town of Churchill

C.5 Universities (Political Science Professors From):

University of Calgary

University of Toronto

Annex D Data Acquisition Software

The following is an account of the commercial software approaches investigated circa Sept 2006. The accuracy of information on commercial products is limited to the accuracy with which the information was provided in consultations with the vendors.

D.1 Importing Logs Into Database for Querying

An early approach considered was to import the log data into Excel® and convert it into an Access database. Both packages are available on DRENET, thus avoiding the administration of obtaining and installing non-standard software. The database could then be queried for the volume of email traffic between pairs of relevant mailboxes. Due to the following factors, this approach was not pursued.

1. Only 65,536 records can be imported into an Excel® spreadsheet. At the time, it was not known that the user base and log files were broken down on a per-server basis. Afterward, however, it was found that the sample log file from a just a single server contained 357,499 physical lines, thus rendering Excel® unusable as an intermediate application.

2. The Exchange Server® 5.5 log files were later found to be variable length, in terms of the number of fields per record, with some records consuming multiple physical lines. Basically, the information for each sender and recipient for an email occupies its own physical line, while an empty physical line terminates the record for the email as a whole (Annex H). Since this data structure does not correspond to a table, it begs the question of whether it makes sense to import the raw log data into a spreadsheet, or even a database. Instead, what is needed is a data preconditioning phase that replaces each log entry for a one-to-many email with several artificial one-to-one emails (along with attenuation weights, as per Section 3.3). Unfortunately, this increases the record count, the maximum of which is already well exceeded.

3. Aside from questionability of treating the raw data as table data, the number of records was too large for a spreadsheet even after accounting for the fact that one record consumes multiple physical lines. The above sample log file for just a single server contained 103,663 multi-line records.

4. At the time that Excel® was being explored, the format of the log files was not known. It would have been discovered upon later examination, however, that determining the sender/recipient IDs would require further nontrivial pre-processing.

D.2 Use of Excel®'s PivotTable® in Multi-National Experiment (MNE) 4

Discussion was held with a social network analyst for MNE 4 to glean lessons learned[49], particularly regarding methodology and tools. The MNE 4 SNA differed considerably in that it examined a variety of communications types. The source data involved records of communications from CFBLNet. Being an experiment-specific network, it stood up for the finite duration of an experiment and has a small overall user base. For MNE 4, there were approximately 800 participants. The source data provided for the SNA encompassed communications between 200 users. It still involved a large volume of data, however, as well as extensive manual preparation so that the data could be read into Excel®. The PivotTable® feature [11] could then generate input suitable for SNA software, which for MNE 4 was UCINET, Pajek, and Cyram's NetMiner.

Even though Polar Guardian would examine email spanning a greater number of topics, it is possible Excel®'s PivotTable® could have been used after the voluminous logs were filtered to leave only records of communication between mailboxes of interest. It seemed more direct, however, to compile the traffic stats using the same scripting step as that used to filter the logs. Such an approach also eliminates the risk due to Excel® limitations in record count. As in the database approach, however, the nontrivial issue of determining the sender/recipient IDs would still need addressing regardless of whether PivotTable® or scripted compilation was used (this wasn't apparent at the time when PivotTable® was being considered).

D.3 Quest®'s MessageStats™

On several occasions, DIMEI suggested using the commercial package MessageStats™ from Quest® to generate reports containing the input data for Pajek (Section 3.4, item 3). They were also quite interested in MessageStats™ as a potential tool for their own use in the long term. A number of factors, however, suggested that this was not the tool to use for the SNA due to the short time frame and limited financial resources.

1. Discussion with Quest® indicates that MessageStats™ needs to query the Active Directory (AD) on the DWAN to identify the sender/recipient. They are of the position that the log files do not contain these details.

2. Despite conversations with several representatives at Quest®, it was unclear whether an output report would be close in format to that required by Pajek. Confirmation was to be provided some time after 18 Sept 2006, and it was still forthcoming when notice about the cessation of project activities was provided to them on 12 Oct 2006.

3. The initial costing scheme yielded an unrealistic cost. The initial cost was $7.50 for each and every mailbox in the domain of interest. As a special consideration, this was then reduced to all mailboxes on all servers of interest. This still results in excessive cost for the final interagency SNA, for which 60 servers was assumed, each serving 1000 mailboxes; 60,000 mailboxes would cost $450,000. According to DIMEI, MessageStats™'s kind of

[49] Hannah State-Davey, Section J.4.

functionality should cost in the order several tens of thousands of dollars, with the absolute upper limit approaching $100K. The only way for the initial cost from Quest® to fall within this ballpark is if the mailbox count was over-estimated by an order of magnitude. Quest® committed to provide more tenable pricing meant for corporate level purchases some time after 18 Sept 2006; it was forthcoming as of 12 Oct 2006.

4. MessageStats™ requires the use of SQL Server, enterprise version (version 2000 preferred). An inquiry was submitted on 18 Sept 2006 to SEAMS about its availability. As of 25 Oct 2006, SEAMS's plans for providing database services consisted of intentions to conduct a study of CFEC's needs. For the SNA in the immediate term, it was not clear from Quest® whether MSDE could be used (free, lighter version of SQL Server); it was not recommended for further exploration, however, due to the large volume of data expected. The size of the database created by MessageStats™ depends primarily on the number of mailboxes hosted by the servers whose log files are processed rather than the number of mailboxes of interest. It was not clear, however, how large the database would be.

Due to DIMEI's recommendation of this application, EXORT decided to reconsider its use when further information was provided, especially with respect to cost and confirmation of functional suitability. This includes the following, many of which were forthcoming as of 12 Oct 2006.

1. Confirmation of suitable output format.

2. Confirmation that generating *N* metrics of traffic volume for *N* mailbox pairs of interest will not require *N* queries.

3. Confirmation that proper compilation of data can be done when the log files come from servers on different domains.

4. An idea of the size of the database generated, and resources required.

5. Demo of its capabilities, which had been proposed by both DIMEI and Quest® a number of times. Quest® has tied the demo to DIMEI's continuing involvement and dialogue, despite DIMEI's request that the demo be driven by the SNA's requirements. EXORT also requested of Quest® that, for the demo, the interest arising from the immediate needs of Polar Guardian be decoupled from DIMEI's interest in meeting future corporate needs. DIMEI has grown silent on the matter of the demo. Considering the cessation of Polar Guardian and the lack of information about functional suitability and cost effectiveness, the MessageStats™ demo was not pursued.

MySQL was mentioned in in-house discussions about the MessageStats™'s requirements for an SQL database. If and when MessageStats™ started to look like a viable alternative, MySQL's suitability could be explored further.

D.4 PROMODAG™ Reports

PROMODAG™ Reports requires access to the DWAN's Active Directory (AD) to identify the users in the tracking logs. It also requires SQL Server. It is not clear how it handles servers from different organizations. PROMODAG™ is another vendor whose position is that the tracking

logs do not contain sufficient information with which users can be identified, and that what data is present takes varying forms.

D.5 Waterford Technologies' MailMeter Insight

There was no response from Waterford Technologies about its suitability for the SNA's requirements.

D.6 Symantec™'s BindView™

According to a representative, this product does not satisfy the SNA's functional requirements. He suggested a scripted data processing approach, such as VB or Perl. Since BindView™ was recently acquired by Symantec™, it could be confusing to establish a POC for product information.

D.7 Morphix's MetaSight®

MetaSight is an SNA analysis package that requires no special pre-processing of log files, since it polls the servers directly. This manner of operation makes it unsuitable for Polar Guardian's SNA, since EXORT only have access to the servers' log files. The costs are also quite high. Simply trying it out for 3 months costs $50K, while full usage costs $180K/year, and additional $50K/server outside of the host organization (DND). SQL Server and IIS (a Microsoft™ web server) are also required.

D.8 Orgnet's InFlow

InFlow performs SNA analysis. It does not read email log files to compile input data. Hence, it fits the same part of the analysis chain as Pajek and requires the same pre-processing being sought.

Annex E Perl References

The following references were vetted based on customer comments on Amazon.com and practitioners' opinions on usenet. The criterion was suitability for experienced programmers. Emphasis on accelerated ramp-up, expedience, and practicality took precedence over elaborations on Perl's philosophical and linguistic elegance.

- Greg London (24 July 2004). *Impatient Perl* (Version: 24 July 2004). [On-line]. Available: www.greglondon.com/iperl [14 Nov. 2006].
- Bradley M. Kuhn (Jan. 2001). *Picking Up Perl* (0.12th ed.). [On-line]. Available: ebb.org/PickingUpPerl [14 Nov. 2006]
- Randal Schwartz, Tom Christiansen, and Larry Wall. Learning Perl (4th ed.). O'Reilly Media, 14 July 2005.
- James Lee. *Beginning Perl* (2nd ed.). Apress, 30 Aug. 2004.
- Tom Christiansen and Nathan Torkington. *Perl Cookbook* (2nd ed.). O'Reilly Media, 21 Aug. 2003.
- Larry Wall, Tom Christiansen, and Jon Orwant. *Programming Perl* (3rd ed.). O'Reilly Media, 14 July 2000.
- Randal L. Schwartz, Tom Phoenix, and Brian D Foy. *Intermediate Perl*. O'Reilly Media, 8 Mar. 2006.
- Allen B. Downey (16 Apr. 2003). *Learning Perl the Hard Way* (Version 0.9). [On-line]. Green Tea Press. Available: greenteapress.com/perl [14 Nov. 2006].
- Tim Maher. *Minimal Perl for UNIX and Linux People*. Manning Publications, 1 Sept. 2005.

Annex F Exchange Server® 5.5 Tracking Log and Events

The following tracking log description was obtained from Microsoft™ support, and is accessible online [17]. Study of example logs indicated some errors in these specifications. For example, field 11 (cost) rarely has value 1, though apparently, it should always be 1. The number of recipients is in field 13 rather than 12. Each recipient name is not tab delimited as implied (it is preceded by a "newline" for each recipient).

F.1 Tracking Log

The tracking log is stored in Exchsrvr\tracking.log. Each day, a new log is created that records one day's activities on the server. Each daily log is named by the date on which it was created, in *yyyymmdd.log* format. The file name date, like all time in the tracking log, is in UTC.

The log can be displayed in any text editor, imported into spreadsheets such as Microsoft Excel, or used as input data to custom applications.

Activities recorded in the tracking log often include a message ID, which is a unique message identifier. By searching the tracking log for the message ID, you can follow the message as it is handled and transported within the site.

The Microsoft Exchange Server Administrator program includes an automated message tracking process. The **Track Message** command traces messages through all existing logs in the network. You can use this process instead of attempting a manual search of the logs.

Interpreting Tracking Log Fields

The following table describes the tab-separated columns in the tracking logs.

Field #	Field Name	Description
1	Message ID or MTS-ID	Message ID is a unique identifier assigned to the message by Microsoft Exchange Server. It stays with the message from its origination to delivery or transfer from the network. Messages from foreign systems include a message transfer system-ID (MTS-ID) that uniquely identifies the component that transported the message.
2	Event #	Represents the event type. For event details, see "Interpreting Events" later in this chapter.

3	Date/Time	Date and time of the event UTC.
4	Gateway name	Name of the gateway or connector that generated the event. If no gateway was involved, the field is blank.
5	Partner name	Name of the messaging service associated with the event. In Microsoft Exchange Server, the partner is the MTA or the information store.
6	Remote ID	Message ID used by the gateway.
7	Originator	Distinguished name of the originating mailbox, if known.
8	Priority	Priority set by the sender. 0 = Normal 1= High -1 = Low
9	Length	Message length in bytes.
10	Seconds	Transport time in seconds. Not used by Microsoft Exchange Server. The value in this field is 0 or blank.
11	Cost	Cost per second for message transfer. Not used by Microsoft Exchange Server. The value in this field is always 1.
12	Recipients	Number of recipients.
13	Recipient name	Distinguished name of the recipient of the message or a proxy address. This field is separated from the previous field by a line feed. This field is repeated for each recipient.
14	Recipient report status	A number representing the result of an attempt to deliver a report to the recipient. Delivered = 0

		Not delivered = 1 This is used only for reports. On other events, it is blank. This field is repeated for each recipient.

F.2 Interpreting Events

The following table defines event numbers that appear in tracking logs.

Event #	Event Type	Description
0	Message transfer in	The MTA completed transfer of responsibility for a message from a gateway, X.400 link, or MTA into the local MTA.
1	Probe transfer in	The MTA completed transfer of responsibility for a probe from a gateway, X.400 link, or MTA into the local MTA.
2	Report transfer in	The MTA completed transfer of responsibility for a report from a gateway, X.400 link, or MTA into the local MTA.
4	Message submission	A message was submitted by a local e-mail client (usually through the information store).
5	Probe submission	An X.400 probe was submitted by a local e-mail client (usually through the information store).
6	Probe transfer out	The MTA completed transfer of responsibility for a probe from the local MTA to a gateway, X.400 link, or another MTA.
7	Message transfer out	The MTA completed transfer of responsibility for a message from the local MTA to a gateway, X.400 link, or another MTA.
8	Report transfer out	The MTA completed transfer of responsibility for a report from the local MTA to a gateway, X.400 link,

		or another MTA.
9	Message delivered	The MTA completed delivery of a message to local recipients (usually through the information store).
10	Report delivered	The MTA completed delivery of a receipt or NDR to local recipients (usually through the information store).
26	Distribution list expansion	The MTA has expanded a distribution list to produce a new message that has recipients who are distribution list members.
28	Message redirected	The MTA has redirected a message or probe to an alternate recipient because of incorrect configuration data for the original recipient, or failure to route the object or reassignment of data contained in the message.
29	Message rerouted	The MTA has rerouted a message, report, or probe because of problems with next route X.400 link or MTA.
31	Downgrading	The MTA has mapped a message, report, or probe into the 1984 X.400 protocol before transferring it to a remote 1984 MTA.
33	Report absorption	The MTA has scheduled a report for deletion because the user did not request it. In X.400 protocol, NDRs are always routed back to the sending MTA even if the user did not request a report.
34	Report generation	The MTA has created a delivery receipt or NDR.
43	Unroutable report discarded	The MTA has discarded a report because the report cannot be routed to its destination.
50	Gateway deleted message	The administrator deleted an X.400 message that was queued by the MTA for transfer to a gateway. No delivery report is generated.

51	Gateway deleted probe	The administrator deleted an X.400 probe that was queued by the MTA for transfer to a gateway. No delivery report is generated.
52	Gateway deleted report	The administrator deleted an X.400 report that was queued by the MTA for transfer to a gateway. No delivery report is generated.
1000	Local Delivery	The sender and recipient are on the same server.
1001	Backbone transfer in	Mail was received from another Messaging Application Programming Interface (MAPI) system across a connector or gateway.
1002	Backbone transfer out	Mail was sent to another MAPI system across a connector or gateway.
1003	Gateway transfer out	The message was sent through a gateway.
1004	Gateway transfer in	The message was received from a gateway.
1005	Gateway report transfer in	A delivery receipt or NDR was received from a gateway.
1006	Gateway report transfer out	A delivery receipt or NDR was sent through a gateway.
1007	Gateway report generation	A gateway generated an NDR for a message.
1010	SMTP Queued Outbound	Outbound mail was queued for delivery by the Internet Mail Service.
1011	SMTP Transferred Outbound	Outbound mail was transferred to an Internet recipient.
1012	SMTP Received Inbound	Inbound mail was received from by the Internet Mail Service.
1013	SMTP Transferred Inbound	Mail received by the Internet Mail Service was transferred to the Information Store.

1014	SMTP Message Rerouted	An Internet message is being rerouted or forwarded to the proper location.
1015	SMTP Report Transferred In	A delivery receipt or NDR was received by the Internet Mail Service.
1016	SMTP Report Transferred Out	A delivery receipt or NDR was sent to the Internet Mail Service.
1017	SMTP Report Generated	A delivery receipt or NDR was created.
1018	SMTP Report Absorbed	The receipt or NDR could not be delivered.

Annex G Exchange Server® 2003 Tracking Log and Reference to Events

The following tracking log description was obtained from DIMEI. Parallel information for Exchange *2000 Server* is available online [18], and it appears to be identical in content to that for Exchange Server® 2003. Information on Event IDs for Exchange Server® 2003 [19] became available during the final revisions of this note.

Exchange 2003 Tracking Logs Fields

The tracking log file is stamped with the following information at the very start of the file.

Message Tracking Log File
Exchange System Attendant Version 6.5.xxxx

The following is a list of all the information available in columnar form in the tracking log file:

Field number	Field name	Description
1	Date	Date of the event.
2	Time	Greenwich mean time of the event.
3	Client-IP	IP of connecting client.
4	Client-hostname	Hostname of connecting client.
5	Partner-name	Name of the messaging service that the message is handed off to. In Exchange 2000, the service can be: SMTP, X400, MAPI, IMAP4, POP3, STORE. This is essentially the same as Exchange Server 5.5, but in Exchange 2000, there are more possibilities for this field.
6	Server-hostname	Hostname of the server that is making the log entry.
7	Server-IP	IP of the server that is making the log entry.
8	Recipient-address	Message recipient (SMTP or X.400 address).
9	Event-ID	Integer corresponding to the Event ID of the action logged, for example: sent, received, delete, retrieve.
10	MSGID	Message ID.
11	Priority	The priority is represented by -1 if low, 0 if normal, 1 if high
12	Recipient-Report-Status	A number representing the result of an attempt to deliver a report to the recipient: 0 if delivered, 1 if not delivered. This is used only for reports (non-delivery reports [NDRs], delivery receipts [DRs]). On other events, it is blank.
13	Total-bytes	Message size (in bytes).

14	Number-recipients	Total number of recipients.
15	Time-taken	Delivery time (in seconds) representing the time it takes to deliver the message. Determined from the difference between the timestamp and time encoded in Message ID. Only valid for messages within the Exchange organization (all versions); there is no requirement to decode other product message IDs such as Sendmail, and so on.
16	Encryption	For the primary body part: 0 if no encryption, 1 if signed only, 2 if encrypted. This is per message, not per recipient.
17	Service-version	Version of the service making the log entry.
18	Linked-MSGID	If there is a MSG ID from another service, it is given here to link the message across services.
19	Message-subject	The subject of the message, truncated to 256 bytes.
20	Sender-address	Primary address of the originating mailbox, if known. This could be SMTP, X.400, or Distinguished Name (DN), depending on transport.

This page intentionally left blank.

Annex H Exchange Server® 5.5 Tracking Log Example Records

The following are selected records from the Exchange Server® 5.5 example tracking log illustrating some of the harder to use identification data. Delimiting tabs are shown as solid right-arrowheads. "Newlines" are shown as scripted backward "P". Sender identity data is underlined. Data for each recipient occupies its own line, immediately following the originator line. Inconsistent use of commas and spacing is preserved, as is upper/lower case. Annex F notes discrepancies between the tracking log data and the specifications.

Anonymization measures are as follows. Square brackets show optional content.

- Actual names have been replaced by generic names, though dashes and " O' " are retained
- Some alphanumeric field values have been replaced with italicized example strings, each of which are optionally appended with integers e.g. *EXAMPLESTRINGn, EXSTRn, XSTRn, SOMECITYn, SOMETOWN, ORGn*, *somenet1*, etc.
- Some decimal digits have been replaced by *D*
- Four-digit years have been replaced with *YYYY*, month/day decimal digits have been replaced by *D*
- Times have been replaced by *H[H]:M[M]:S[S]*, where the number of digits have been preserved.

C=CA;A=*ORG4.COUNTRY*;P=*ORG3*;L=*EXAMPLESTRING5*►0►*YYYY.D.DD H:M:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING1*/CN=*SOMECORP1*
MTA►►C=CA;A=*ORG4.COUNTRY*;P=*ORG3*;DDA:SMTP=*EmailUserName1*(a)*somenet1.org1.ca;*►0►8214►0►0►►3¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=Recipients/cn=+*XMPLSTRING1*¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=Recipients/cn=*LASTNAME1, FIRSTNAME1* 389¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=Recipients/cn=*LASTNAME2, FIRSTNAME2* 097¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING2*►4►*YYYY.D.DD H:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING4*/CN=*SOMECORP1* PRIVATE
MDB►►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME3, FIRSTNAME3*@931►0►1550►0►0►►1¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;O=*XSTR1*;DDA:SMTP=*EmailUserName2*(a)examplestring26.com;¶
¶
c=CA;a=*ORG4.COUNTRY*;p=*ORG1.ORG2*;l=*EXAMPLESTRING3*►1000►*YYYY.D.DD HH:MM:SS*►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=Configuration/cn=Servers/cn=*EXAMPLESTRING4*/cn=*Somecorp1* Private
MDB►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=Configuration/cn=Servers/cn=*EXAMPLESTRING4*/cn=*Somecorp1* Private

MDB►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=RECIPIENTS/CN=*LASTNAME4, FIRSTNAME4* @947►0►1712►0►1►►2¶
/O=*ORG1.ORG2*/OU=*XSTR1*/CN=RECIPIENTS/CN=+*XSTR1* 190005¶
/O=*ORG1.ORG2*/OU=*XSTR1*/CN=RECIPIENTS/CN=*Lastname31, Firstname31* 446¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING6*►8►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING1*/CN=*SOMECORP1*
MTA►C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING7*►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=?*ORG1 EXAMPLE TEXT NOT REAL*
►1►3596►0►0►►7¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME5, FIRSTNAME5* 276►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME6, FIRSTNAME6* 771►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME7, FIRSTNAME7* 062►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME8, FIRSTNAME8* 346►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME9, FIRSTNAME9* @473►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME10, FIRSTNAME10* 590►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME19, FIRSTNAME19* 964►1¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING14*►0►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING1*/CN=*SOMECORP1*
MTA►►/o=*ORG1.ORG2*/ou=ALERT/cn=RECIPIENTS/cn=XSTRNG1►0►36093►0►0►►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME11, FIRSTNAME11* 509¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING15*►0►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING1*/CN=*SOMECORP1*
MTA►►/o=*ORG1.ORG2*/ou=*SOMECITY4*/cn=RECIPIENTS/cn=O'*LASTNAME12* CAPT MM@*XMPLSTRING1*►1►23140►0►0►►2¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME13, FIRSTNAME13* 793¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME14, FIRSTNAME14* S847¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING8*►7►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING1*/CN=*SOMECORP1*
MTA►►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME15, FIRST-NAME15* 655►0►1574►0►0►►1¶
/o=*ORG1.ORG2*/ou=*SOMECITY1*/cn=RECIPIENTS/cn=*LASTNAME16,FIRSTNAME16* 192¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING16*►9►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING4*/CN=*SOMECORP1* PRIVATE MDB►►/o=*ORG1*-ORG2/ou=*SOME ADMINISTRATIVE GROUP*/cn=RECIPIENTS/cn=*LASTNAME17.INITIALS17*►0►3144►0►0►►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME18, FIRSTNAME18* 985¶
¶

C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING9*►4►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING4*/CN=*SOMECORP1* PRIVATE
MDB►►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME20, FIRSTNAME20* 996►0►8012►0►0►►6¶
/o=*ORG1.ORG2*/ou=*EXSTRING1*/cn=*Some Users*/cn=*lastname21.initials21@org3*.gc.ca¶
/o=*ORG1.ORG2*/ou=*EXSTRING1*/cn=*Some Users*/cn=*lastname22.initials22@org3*.gc.ca¶
/o=*ORG1.ORG2*/ou=*EXSTRING1*/cn=*Some Users*/cn=*lastname23.initials23@org3*.gc.ca¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME24, FIRSTNAME24* 606¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME25, FIRSTNAME25* 016¶
/o=*ORG1.ORG2*/ou=*EXSTRING1*/cn=*Some Users*/cn=*lastname26.initial26@org3*.gc.ca¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING17*►9►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING4*/CN=*SOMECORP1* PRIVATE
MDB►►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=SOME RECIPIENTS/cn=*LASTNAME27, FIRSTNAME27@*524►0►1978►0►0►►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME28, FIRSTNAME28* 237¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING10*►4►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING4*/CN=*SOMECORP1* PRIVATE
MDB►►C=CA;A=*ORG4.COUNTRY*;P=*ORG3*;S=*Surname29*;G=*GivenName29*;I=*INITIALS29*;►0►8093►0►0►►2¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME34, FIRSTNAME34* 874¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME35, FIRSTNAME35* 161¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING18*►0►*YYYY.D.DD HH:MM:S*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING24*/CN=*SOMECORP1*
MTA►►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=123456789►0►3497►0►0►►2¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME36, FIRSTNAME36* 699¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME37, FIRSTNAME37* 229¶
¶
c=CA;a=*ORG4.COUNTRY*;p=*ORG1.ORG2*;l=*EXAMPLESTRING11*►1000►*YYYY.D.DD HH:M:SS*►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=Configuration/cn=Servers/cn=*EXAMPLESTRING4*/cn=*Somecorp1* Private
MDB►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=Configuration/cn=Servers/cn=*EXAMPLESTRING4*/cn=*Somecorp1* Private
MDB►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=*EXAMPLESTRING25*/CN=+*EXSTRING2*►0►901►0►1►►1¶
/O=*ORG1.ORG2*/OU=*XSTR1*/CN=RECIPIENTS/CN=*LASTNAME38, FIRSTNAME38* @835¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING19*►9►*YYYY.D.DD HH:M:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING4*/CN=*SOMECORP1* PRIVATE
MDB►►/o=*ORG1.ORG2*/ou=*SOMECITY5*/cn=RECIPIENTS/cn=*LASTNAME39,*

FIRSTNAME39, 376►0►42274►0►0►►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME40, FIRSTNAME40* 191¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING20*►9►*YYYY.D.DD HH:MM:S*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING4*/CN=*SOMECORP1* PRIVATE
MDB►►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*EXSTRING3*►0►41982►0►0►►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*EXSTR1* @419¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING12*►4►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING4*/CN=*SOMECORP1* PRIVATE
MDB►►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME41, FIRSTNAME41* 918►0►1538►0►0►►1¶
/o=*ORG1.ORG2*/ou=SOMETOWN/cn=RECIPIENTS/cn=+*XSTR2*¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING21*►0►*YYYY.D.DD HH:MM:SS*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING27*/CN=*SOMECORP1*
MTA►►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME42.INITIALS42*►0►1932►0►0►►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME43, FIRSTNAME43* 416¶
¶
C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;L=*EXAMPLESTRING22*►0►*YYYY.D.DD HH:MM:S*►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=CONFIGURATION/CN=SERVERS/CN=*EXAMPLESTRING1*/CN=*SOMECORP1*
MTA►►/o=*ORG1.ORG2*/ou=*SOMECITY1*/cn=RECIPIENTS/cn=*LASTNAME FIRSTNAME,* 642►0►3574►0►0►►1¶
/o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME44, FIRSTNAME44* 933¶
¶
c=CA;a=*ORG4.COUNTRY*;p=*ORG1.ORG2*;l=*EXAMPLESTRING13*►1000►*YYYY.D.DD HH:MM:SS*►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=Configuration/cn=Servers/cn=*EXAMPLESTRING4*/cn=*Somecorp1* Private
MDB►/o=*ORG1.ORG2*/ou=*XSTR1*/cn=Configuration/cn=Servers/cn=*EXAMPLESTRING4*/cn=*Somecorp1* Private
MDB►►/O=*ORG1.ORG2*/OU=*XSTR1*/CN=RECIPIENTS/CN=*LASTNAME45, FIRSTNAME45*179►0►1808►0►1►►1¶
/O=*ORG1.ORG2*/OU=*XSTR1*/CN=RECIPIENTS/CN=*LASTNAME46, FIRSTNAME46* 813¶
¶

Annex I Characteristics of Identification Data in Exchange Server 5.5 Tracking Logs

I.1 Message ID Field

An example of the X.400-like message ID is:

C=ca;A=*org4.country*;P=*org1.org2*;L=*EXAMPLESTRING23*

A brief search reveals likely interpretations for the subfield names

C=country, A=ADMD, P=PrvID

I.2 Sender and Recipient Fields

1. Contain possibly multiple data items, presumably about the same person, separated by semicolon

2. Sometimes contains X.400-like data. For example:

   (i) C=CA;A=*ORG4.COUNTRY*;P=*ORG3*

   (ii) C=CA;A=*ORG4.COUNTRY*;P=*ORG3*;S=Smith;G=Jane;I=JB;

   A brief search reveals likely interpretations for the subfield names: S=surname, G=given name, I=initials (typically, first initial matches given name).

3. Sometimes contains X.400/500-like data. X.500 is a standard for directory services and supports X.400. Examples of X.500-like data in the 5.5 log file:

   (iii) /o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*LASTNAME, FIRSTNAME* 906

   (iv) /o=*ORG1.ORG2*/ou=*SOMECITY2*/cn=RECIPIENTS/cn=*LASTNAME, FIRSTNAME* 433

   (v) /o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=+*ORG6*016016

   A brief search reveals likely interpretations for the subfield names: o=organization, ou=organizational unit (there can be more than one), cn=common name (possibly more than one).

   A further search reveals that these subfield names are often associated with X.400, but when the GAL is accessed from Outlook, they are tagged as X.500 data. For the purpose of distinguishing them from X.400 data above, they are referred to as X.500 data here.

   The final 3-digit numbers in the above examples do not uniquely identify the users, and there did not seem to be any initials for middle names to help disambiguate user identity.

4. Sometimes contains conventional SMTP email address:

   (vi) DDA:SMTP=*Smith.John*(a)ca.*somecorp2*.com

   (vii) DDA:SMTP=*Smith.JW*(a)*org3*.gc.ca

I.3 Outgoing Email Records

From examining outgoing (event#4) records in the example log file, the "Originator" field seems to be exclusively of the form (iii) and (iv) above; both of these are merely different examples of the same form.

The recipients, however, are of the forms (iii,iv), (v), and the following less frequent forms.

1. /o=*ORG1.ORG2*/ou=*EXSTR2 EXSTRING4*/cn=RECIPIENTS/cn=*DUMMYSTRING@EXSTRING4*

2. /o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*EXStr3*, *Example Random String* 2@101

3. /o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=*Xstr3*

4. /o=*ORG1.ORG2*/ou=*XSTR1*/cn=RECIPIENTS/cn=SMITH, JOHN @240

5. /o=*ORG1.ORG2*/ou=*EXSTRING1*/cn=*XMPL RNDM STR*/cn=SMITH.J7@*ORG3*.GC.CA

6. /o=*ORG1.ORG2*/ou=*SomeCity3*/cn=Recipients/cn=Smith.JW

7. C=CA;A=*ORG4.COUNTRY*;P=*ORG1.ORG2*;O=*XSTR1*;DDA:SMTP='Smith(a)state.gov';

In the #1 above, it was not clear whether the dummy string corresponded to a name, user, mailbox, or something else.

Records in forms (iii,iv) obviously correspond to user mailboxes, and identification can be made, with some ambiguity. For some of the forms 1-7 above, it is also obvious that the recipient is a user or a mailbox e.g. #4-7. For others, it is not clear. Of particular concern, however, is the fact that records of form (v) are quite cryptic, and more research or consultation with subject matter experts (SMEs) is required to determine whether identification can be made for all outgoing records.

I.4 Incoming Email Records

From examining event #9 records in the example log file, the approximate reverse of the outgoing email records was observed. The "Originator" field seems to consist of the multitude of formats, as was seen in the recipients of outgoing records. In contrast, the recipients of incoming email records consist of very few formats:

- A significant portion in format (v)

- The bulk of records in format (iii,v)
- Variations of (iii,v), where the numerical component at the end is prefixed and/or suffixed with an letter, and/or sometimes prefixed with "@" or "*"

The same considerations as those for outgoing records apply i.e. recipients can mostly be identified, with some ambiguity, whereas senders are plagued by a plethora of forms of identification data with varying degrees of decipherability.

Annex J Contacts

J.1 DND Personnel

- Mike (Michel) Manor
  - POC for DIMEI 3-6-3 (formerly 3-6-4) regarding acquisition of tracking log files
- Maj Mohammad Chaudhary
  DIMEI 3-6
  - Authority for release of tracking log files
- Kwok-Fai Ha
  Contractor
  DIMEI 3-6-4-C
  - Seemingly the person most knowledgeable in Exchange Server® encountered in-house
- Donald Messier (Major, retired)
  - Participated in 2004 study by DIMEI 3-4 of server-to-server traffic using tracking logs. (All of DIMEI has since consolidated into a single organization, DIMEI).
  - Currently in DWAN GAL as contractor in DIMEI 7
  - OPI who did the work:
    Cherif Djerboua
    Microsoft™
    cherif.djerboua@microsoft.com
    - No longer with DND full time
- Sgt Charles D. Dechamp
  764 Comm. Sqn. DEMS Supv.
  - Before Remi Lagace transferred from SEAMS to 76 Comm, he suggested contacting Sgt Dechamp for information on how JTFN and Canada Command personnel were distributed among the servers
  - Provided the links for downloadable GAL data
- Brian Woolsey
  DIMEI 2 (formerly DDCEI 2)
  Tel: 613-944-4712 / Cell: 613-220-0610
  Program Manager, Software Integration Engineering
  Defence Software Baseline & Life Cycle Product Management
  - Authority for accessing Microsoft™ technical support

J.2 Microsoft™ Support

- Pierre Major, MCSE
  Technical Account Manager
  Microsoft™ Premier Support
  Email: Pmajor@Microsoft.Com
  Phone: 613-232-6606
  Cell: 613-298-4582
    - Initial POC
- Amy-Leigh B. Mack
  Exchange Support Professional
  Microsoft™ Enterprise Messaging Support
  Email: amyma@microsoft.com
  Phone: 980-776-8307
  Office hours: Mon.-Fri. 8am-5pm EST
    - Next POC
- Christopher Nguyen
  Microsoft™ Enterprise Business Application
  Messaging
  E-mail: v-11chng@mssupport.microsoft.com
  Phone: 416-246-5580 ext. 5471
  Hours: Mon.-Fri. 9am-6pm EST
    - Pierre Major's co-facilitator for meeting with "Mark", a former engineer with exposure to development of Exchange Server® 5.5 tracking logs.

J.3 Quest® Software

- Jill Kaser
  Microsoft™ Exchange Specialist
  Jill.Kaser@quest.com
  800-263-0036 ext. 4726
  614-726-4726 direct
    - Main POC for Quest®.
- Rob Sargent
  Head, product team (Product manager)
  Kanata
  613-270-1500
    - Technically knowledgeable
- Eric Hibar
  System Consultant
  Columbus, Ohio
    - Technically knowledgeable

- Pam Turenne
  Sales account manager for DND
  Ottawa

J.4 SNA for Multi-National Experiment (MNE) 4

- Hannah State-Davey
  HMSDAVEY@qinetiq.com
  - Social network analyst for MNE 4
- Mark Round
  MDROUND@qinetiq.com
  - Social network analyst for MNE 4
- Neil G. Verrall
  NGVERRALL@mail.dstl.gov.uk
  - U.K. lead analyst for MNE 4

Bibliography

SNA Books

The following books were subjectively selected and ranked in order of decreasing preference based on customer reviews available on Amazon.com, circa summer 2006. The criteria were (1) positive reviews, (2) emphasis on application of concepts and bringing forth practical relevance rather than elaboration on the mathematical and graph theoretic details behind the metrics, and (3) how many people bought the book after reviewing it. The motivation was to identify what resource might enable a newcomer to ramp up to a working knowledge of the method as quickly as possible, for application rather than furthering the underlying theory. There is no guarantee, however, that the ranking accurately reflects the effectiveness of the resource in meeting this aim. At the time of writing, unfortunately, there seemed to be a shortage of books focusing on the practice rather than the theory.

For this project, Pajek was the SNA tool selected. Therefore, #5 subjectively moves up to #3 for that effort.

These books were collected for ramp-up during the project (either purchased for CORA or borrowed via interlibrary loan), but the investigators primarily studied #1 and #6 during the data collection planning.

1. Cross, R. and Parker, A. (2004), The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations. Boston, MA: Harvard Business School Press.
   - Focus on cases and design of SNA study instead of theory.
   - Best reviews at Amazon: 5 stars, 5 reviewers, 75% purchase rate among viewers
   - Most applied of all books investigated
2. Scott, J.P. (2000), Social Network Analysis: A Handbook, 2nd ed. London, Thousand Oaks, New Delhi: SAGE Publications.
   - Covers theory, metrics, and computer programs
   - Amazon: 5 stars, 1 reviewer, 66% purchase rate among viewers
   - Described as a theoretical context for Wasserman & Faust
3. Wasserman, S., Faust, K., Iacobucci, D., and Granovetter, M. (1994), Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences series). Cambridge, New York, Melbourne: Cambridge University Press.
   - Strong on the graph theory
   - Amazon: 4 stars, 4 reviewers, 55% purchase rate among viewers
4. Carrington, P.J. (Ed.), Scott, J. (Ed.), and Wasserman S. (Ed.) (2005), Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences series). Cambridge: Cambridge University Press.

- Leans towards theory.
- Winner of the 2006 Harrison White Outstanding Book Award from the Mathematical Sociology Section of the American Sociological Association
- Positive review at SAGE Publications, Prof. John Levi Martin, University of Wisconsin, Madison
- Positive review at Canadian Journal of Sociology Online, Prof. Bonnie Erickson, University of Toronto
  - Not for beginners, heavy on theory, presumes Wasserman & Faust as foundation
- Amazon: 55% purchase rate among viewers

5. de Nooy, W., Mrvar, A., and Batagelj, V. (2005), Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences series). Cambridge, New York, Melbourne: Cambridge University Press.
   - Chooses Pajek as the tool for illustration of practice.
   - Amazon: 3.5 stars, 3 reviewers, 19% purchase rate among viewers (69% circa Feb 07)

6. Degenne, A. and Forse, M. (1999), Introducing Social Networks (Introducing Statistical Methods series). London: SAGE Publications.
   - Described as focusing on application rather than theory, but found by EXORT to be theoretical
   - Amazon: 4 star, 1 reviewer, 7% purchase rate among viewers (33% circa Feb 07)

SNA Articles

The following is a suggested list of articles for familiarization with applied SNA. The articles below have been selected for their strong likelihood of being oriented toward application rather than the underlying graph theory. Interpretation/visualization is a major factor, as is a case-study nature. Military focus also weighs into the selection. Most were reviewed to varying degrees.

- Borgatti, S.P (2005), Centrality and Network Flow, *Social Networks: An International Journal of Structural Analysis,* 27 (1), 55-71, Elsevier.
  Available: http://www.analytictech.com/borgatti/papers/centflow.pdf, (Access date: 1 Mar. 2008).
- Chan, K. and Liebowitz, J. (2006), The Synergy of Social Network Analysis and Knowledge Mapping: A Case Study, *Int. J. Management and Decision Making*, 7 (1), 19-35.
  Available: http://www.knowledgeboard.com/download/2787/8169.pdf, (Access date: 1 Mar. 2008).
- Clauset, A., Newman, M.E.J., and Moore, C. (2004), Finding Community Structure in Very Large Networks, *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics* (electronic journal) 70, .no. 6, ID 066111 (6 pages).
  http://scitation.aip.org/dbt/dbt.jsp?KEY=PLEEE8&Volume=70, (Access date: 1 Mar. 2008).

- Coulon, F. (2005), The Use of Social Network Analysis in Innovation Research: A Literature Review, *DRUID Academy Winter PhD Conference*.
  Available: http://www.druid.dk/uploads/tx_picturedb/dw2005-1613.pdf, (Access date: 1 Mar. 2008).
- Dekker, A. (2005), Conceptual Distance in Social Network Analysis, *Journal of Social Structure* (electronic journal), 6 (3).
  https://www.cmu.edu/joss/content/articles/volume6/dekker, (Access date: 1 Mar. 2008).
- Dekker, A. (2001), Visualisation of Social Networks Using CAVALIER, In *Proc. 2001 Asia-Pacific Symposium on Information Visualisation*, 9, 49-55, Darlinghurst, Australia, Australia: Australian Computer Society, Inc.
  Available: http://crpit.com/confpapers/CRPITV9Dekker.pdf, (Access date: 1 Mar. 2008).
- Dekker, A. (2000), Social Network Analysis in Military Headquarters Using CAVALIER, In *Proc. 5th International Command and Control Research and Technology Symposium,* Command and Control Research Program.
  Available: http://citeseer.ist.psu.edu/anthony00social.html (Access date:14 Nov. 2006).
  Available: http://www.dodccrp.org/events/5th_ICCRTS/papers/Track6/039.pdf, (Access date: 1 Mar. 2008).
  Available: http://www.insna.org/Connections-Web/Volume24-3/Dekker.web.pdf, (Access date: 1 Mar. 2008).
- Dight, C. (2 Mar. 2006), Profit Potential Laid Bare in E-Mail Links, *Times Online* (electronic news).
  http://www.timesonline.co.uk/article/0,,8171-2063077,00.html, (Access date: 1 Mar. 2008).
- Huisman, M. and van Duijn, M.A.J (2005), Software for Social Network Analysis, In book #4 of this bibliography, pp. 270-316.
- McGregor, J. (2006), The Office Chart That Really Counts, *BusinessWeek Online* (electronic news).
  http://www.businessweek.com/magazine/content/06_09/b3973083.htm, (Access date: 1 Mar. 2008).
- Pearson, M. and West, P. (2003), Drifting Smoke Rings: Social Network Analysis and Markov Processes in a Longitudinal Study of Friendship Groups and Risk-Taking, *Connections: Official Journal of International Network for Social Network Analysis*, 25 (2), 59-76.
  Available: http://www.insna.org/Connections-Web/Volume25-2/3.Pearson.pdf, (Access date: 1 Mar. 2008).
- San Martin, M. and Gutierrez, C. (2006), A Database Perspective of Social Network Analysis Data Processing, *Sunbelt XXVI International Sunbelt Social Network Conference*.
  Available: http://www.dcc.uchile.cl/~cgutierr/ftp/sunbelt.pdf, (Access date: 1 Mar. 2008).
- Yang, S. and Knoke, D. (2001), Optimal Connections: Strength and Distance in Valued Graphs, *Social Networks: An International Journal of Structural Analysis*, 23 (4), 285-295, Elsevier.
  Available: http://comp.uark.edu/~yangwang/yang2001.pdf, (Access date: 14 Nov. 2006).
  Available: http://www.soc.umn.edu/~yang/resume/article.pdf, (Access date: 1 Mar. 2008).

Available: http://www.soc.umn.edu/~knoke/pages/Yang&Knoke.pdf, (Access date: 1 Mar. 2008).

List of symbols/abbreviations/acronyms/initialisms

AD	Active Directory
ASIWG	Arctic Surveillance Interdepartmental Working Group
C&S	Command and Sense (team)
CFBLNet	Combined Federated Battle Lab Network
CFEC	Canadian Forces Experimentation Centre
CFWC	Canadian Forces Warfare Centre
CSV	Comma separated values
DIMEI	Director Information Management Engineering and Integration
DND	Department of National Defence
DRDC	Defence Research and Development Canada
DRDKIM	Director Research and Development Knowledge and Information Management
DWAN	Defence Wide Area Network
EXORT	Experimentation Operational Research Team
GAL	Global Address List
GB	Gigabyte
ID	Identity
IIS	Microsoft™ Internet Information Services
IP	Internet Protocol (address)
IT	Information Technology
JTFN	Joint Task Force North
K	Kilo (1000)
MB	Megabyte
MNE	Multi-National Experiment
MSDE	Microsoft™ Data Engine Microsoft™ Desktop Engine Microsoft™ SQL Server Desktop Engine
MTA	Mail Transfer Agent
MySQL	An SQL Database Management System
OGD	Other Government Department
OR	Operational Research
POC	Point of Contact

POP3	Post Office Protocol version 3
PSEPC	Public Safety and Emergency Preparedness Canada
RCMP	Royal Canadian Mounted Police
RFC	Request For Comment
RFC2822	Internet Engineering Task Force RFC document defining the format of SMTP email
S&T	Science and Technology
SEAMS	Synthetic Environment and Modelling & Simulation
SME	Subject Matter Expert
SMTP	Simple Mail Transfer Protocol
SNA	Social Network Analysis
SOP	Standard Operating Procedure
SQL	Structured Query Language
VB	Microsoft™ Visual Basic
X.400	Message exchange standard
X.500	Series of computer networking standards for electronic directory services

Distribution list

Document No.: DRDC CORA TM 2009-030

LIST PART 1: Internal Distribution by Centre:

DRDC CORA
1 DG CORA
1 EXORT
2 CORA Library (1 hard copy, 1 CD)
6 Authors (2 copies each)

10 TOTAL LIST PART 1 (9 hard copies, 1 CD)

LIST PART 2: External Distribution by DRDKIM

1 DRDKIM (1 CD)
1 Canadian Forces Experimentation Centre (CFEC)
1 DGIP (Attention: Stephan Flemming)
2 DRDC Toronto (Attention: Dr Renee Chow, Dr David Smith)

5 TOTAL LIST PART 2 (4 hard copies, 1 CD)

15 TOTAL COPIES REQUIRED (13 hard copies, 2 CDs)

This page intentionally left blank.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

DRDC -- Centre for Operational Research and Analysis
National Defence Headquarters, 101 Colonel By Drive
Ottawa, Ontario K1A 0K2

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C, R or U) in parentheses after the title.)

Data Acquisition and Preparation for Social Network Analysis Based on Email: Lessons Learned

4. AUTHORS (last name, followed by initials – ranks, titles, etc. not to be used)

Ma, F.; Allen, D.; Dooley, P.

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
June 2009	99	19

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Technical Memorandum

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

Canadian Forces Experimentation Centre
Building 4, Shirley's Bay Campus
101 Colonel By Drive
Ottawa, ON K1A 0K2

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)
N/A	N/A

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC CORA TM 2009-030	N/A

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

(X) Unlimited distribution
( ) Defence departments and defence contractors; further distribution only as approved
( ) Defence departments and Canadian defence contractors; further distribution only as approved
( ) Government departments and agencies; further distribution only as approved
( ) Defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

See Distribution List

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

In sharing information to improve situational awareness, other government departments and remotely situated outposts may vary in their reporting of information. A social network analysis was initiated within the Department of National Defence to show where informal communication may be significant to information sharing. The study was undertaken circa Q3 2006 by the Experimentation Operational Research Team at the Canadian Forces Experimentation Centre for the Command and Sense Team. Analytical results are not available, as the undertaking was not completed. This report describes the lessons learned in planning the data collection and preparation for the social network analysis.

The work was done under project Polar Guardian, the goal of which was to assess situational awareness in the arctic. The plan for the social network analysis included an initial email-based phase and a follow-up survey-based phase. This report focuses on the email phase; it is not a comparison of the two phases as separate approaches.

Due to the short time frame for conducting the trial on the social network analysis approach, in-house methods for data acquisition and analysis were explored. The main challenges in this approach arise from generating the communications data from email tracking logs in isolation from other information gathering and information providing parts of a corporate computer network.

Commercial tools were investigated, and warrant further examination. Their use requires a longer time frame for approval and installation on the Defence Wide Area Network.

Of the commercial and home-grown approaches, the most time is likely needed for solutions involving access to the servers, and deployment of applications on them.

Direct access to subject matter expertise in email administration is essential to arriving at a means for effective and timely data gathering and preparation. Such access is also essential for an interagency social network analysis, the issues of which are touched upon in this technical memorandum only at a high level.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Social Network Analysis (SNA); Situation Awareness (SA); Arctic; Information Sharing; Interagency; Polar Guardian; Email; Exchange Server

DRDC CORA

www.drdc-rddc.gc.ca