Defence Research and Development Canada
Recherche et développement pour la défense Canada

DEFENCE R&D DÉFENSE

Applicability of Social Network Analyses in an Intelligence Context

Etienne Martineau
DRDC Valcartier

Defence R&D Canada – Valcartier

Technical Memorandum
DRDC Valcartier TM 2013-418
December 2012

Canada

Applicability of Social Network Analyses in an Intelligence Context

Etienne Martineau
DRDC Valcartier

Defence R&D Canada – Valcartier

Technical Memorandum
DRDC Valcartier TM 2013-418
December 2012

© Her Majesty the Queen in Right of Canada, as represented by the Minister of National Defence, 2012

© Sa Majesté la Reine (en droit du Canada), telle que représentée par le ministre de la Défense nationale, 2012

Abstract

This document presents a classification of theoretical social network analyses to determine if they are suitable for the military intelligence context. This work was done under the project 15do, "Social network analysis in a counter insurgency context", to orient the development of a proof of context prototype. Thirty-four (34) theoretical social network analyses were selected and classified based on the implementation risk they pose. The evaluation criteria used are computational complexity, topological and flow requirements and missing data robustness. The risk tolerance of the military intelligence process was also evaluated using the same criteria. Using both results, it was possible to identify 8 suitable algorithms, 11 unusable and 15 risky ones. These results suggest that classical social network analysis methods are not sufficient to answer the military requirements stated in this project. Alternative ways to extract information and perform network analysis should be investigated. In particular, there is a need for algorithms that are robust to missing data and can extract sub-networks of interest efficiently. The impact of missing data is still not well understood and should be the focus of further research as it is largely encountered in the context of covert networks.

Résumé

Ce document présente une classification des analyses des réseaux sociaux pour déterminer si elles sont utilisables dans le contexte du renseignement militaire. Ce travail a été effectué sous le projet 15do, « Social network analysis in a counter insurgency context », pour orienter le développement d'une preuve de concept logiciel. Trente-quatre (34) analyses de réseaux sociaux théoriques ont été sélectionnées et classées, basé sur le risque d'implémentation qu'elles posent. Les critères d'évaluation utilisés sont : la complexité algorithmique, les requis topologiques et de flux et la robustesse aux valeurs manquantes. La tolérance au risque du processus du renseignement militaire a été aussi évaluée à l'aide des mêmes critères. En utilisant ces deux résultats, il est possible d'identifier 8 algorithmes utilisables, 11 non-utilisables et 15 à haut risque. Ces résultats suggèrent que les analyses de réseaux sociaux classiques ne suffisent pas à répondre aux besoins des militaires énoncés dans ce projet. Des méthodes alternatives d'extraction d'information et d'analyse de réseaux devraient être recherchées. En particulier, il existe un besoin pour des algorithmes robustes aux données manquantes et qui peuvent extraire des sous-réseaux d'intérêt efficacement. De plus, l'impact des données manquantes n'est pas bien compris et devrait faire l'objet de recherches plus poussées puisque c'est un phénomène souvent rencontré dans le contexte des réseaux secrets.

This page intentionally left blank.

Executive summary

Applicability of Social Network Analyses in an Intelligence Context

Etienne Martineau; DRDC Valcartier TM 2013-418; Defence R&D Canada – Valcartier; December 2012.

Introduction or background: This Technical Memorandum documents results and findings from research activities conducted at Defence R&D Canada under Project 15do. This project, entitled "Social network analysis in a counter insurgency context" is part of DRDC's Applied Research Program. The main objective of this project is to investigate social network analysis techniques, methodologies and tools in order to enhance the Canadian Forces intelligence analysis capability and enable sense making based on analysis results in a counter insurgency context.

Results: This document presents a classification of theoretical social network analyses to determine if they are suitable for use in the military intelligence context. To do so, 34 theoretical social network algorithms were selected and classified based on the implementation risk they pose. The criteria used for the evaluation are computational complexity, topological and flow requirements and missing data robustness. The risk tolerance of the military intelligence process was also evaluated using the same criteria. Building on both results, it was possible to identify 8 suitable algorithms, 15 risky ones and 11 unusable.

Significance: The results show that some theoretical social network analyses are not directly applicable to the military intelligence process. Some analyses are robust and can provide useful and timely information to the analyst. However, most of them are tools that are meant to be used in a controlled environment that the military context cannot provide.

Future plans: These results suggest that classical social network analysis alone is not sufficient, as it stands today, to answer the needs of this project. Alternative ways to extract information and perform network analysis should be investigated. In particular, there is a need for algorithms that are robust to missing data and that can extract sub-networks of interest efficiently. Also, missing data is not well understood and should be the focus of further research.

Sommaire

Applicability of Social Network Analyses in an Intelligence Context

Etienne Martineau ; DRDC Valcartier TM 2013-418 ; R & D pour la défense Canada – Valcartier; décembre 2012.

Introduction ou contexte : Ce mémorandum technique documente les résultats d'activités de recherche effectuées à Défense R&D Canada sous le projet 15do. Le projet "Social network analysis in a counter insurgency context" fait partie du programme de recherche appliqué de RDDC. L'objectif principal de ce projet est d'investiguer les techniques, méthodologies et outils d'analyse de réseaux sociaux dans le but d'augmenter la capacité d'analyse et d'éveil situationnel en renseignement des Forces Canadiennes, basé sur les résultats d'analyse dans un contexte de contre-insurrection.

Résultats : Ce document présente une classification des analyses des réseaux sociaux pour déterminer si elles sont utilisables dans le contexte du renseignement militaire. Pour ce faire, trente-quatre (34) analyses de réseaux sociaux théoriques ont été sélectionnées et classées, basé sur le risque d'implémentation qu'elles posent. Les critères d'évaluations utilisés sont : la complexité algorithmique, les requis topologiques et de flux et la robustesse aux valeurs manquantes. La tolérance au risque du processus du renseignement militaire a été aussi évaluée à l'aide des mêmes critères. En utilisant ces deux résultats, il est possible d'identifier 8 algorithmes utilisables, 11 non-utilisables et 15 à haut risque.

Importance : Les résultats montrent que quelques analyses de réseaux sociaux théoriques ne sont pas directement applicables au processus de renseignement militaire. Quelques analyses sont robustes et peuvent fournir de l'information pertinente à l'analyste en temps raisonnable. Par contre, la plupart sont des outils qui sont conçus pour être utilisés dans des environnements contrôlés que le contexte militaire ne peut pas fournir.

Perspectives : Ces résultats suggèrent que l'analyse de réseaux sociaux classique à elle seule n'est pas suffisante pour répondre au besoin de ce projet. Des façons alternatives d'extraire de l'information et d'effectuer l'analyse des réseaux sociaux devraient être explorées. En particulier, il existe un besoin pour des algorithmes robustes aux données manquantes et qui peuvent extraire des sous-réseaux d'intérêt efficacement. De plus, l'impact des données manquantes n'est pas bien compris et devrait faire l'objet de recherches plus poussées.

Table of contents

Abstract ……. i
Résumé ……. i
Executive summary iii
Sommaire iv
Table of contents v
List of figures vii
List of tables viii
1 Introduction 1
2 Categorisation of SNA techniques 3
2.1 Evaluation method and criteria 3
2.1.1 Computational complexity 3
2.1.2 Topological and Flow requirement 4
2.1.3 Missing data robustness 5
2.2 SNA algorithm assessment 6
2.2.1 Centrality measures 6
2.2.1.1 Degree centrality 7
2.2.1.2 Eigenvector centrality 7
2.2.1.3 Freeman closeness 7
2.2.1.4 Topological centrality 7
2.2.1.5 Freeman Betweenness 8
2.2.1.6 Bridging centrality 8
2.2.1.7 Flow Betweenness 8
2.2.1.8 Dependence centrality 8
2.2.1.9 Position role centrality 9
2.2.1.10 Newman Random walk centrality 9
2.2.2 Other node measures 9
2.2.2.1 Structural equivalence 9
2.2.2.2 Automorphic and Isomorphic equivalence 10
2.2.2.3 Regular equivalence 10
2.2.2.4 Local role equivalence 10
2.2.2.5 Brokerage 11
2.2.2.6 Clustering coefficient 11
2.2.3 Node partitioning algorithm 11
2.2.3.1 CONCOR 11
2.2.3.2 Hierarchical clustering 11
2.2.3.3 Blockmodels 12
2.2.4 Graph measures 12

2.2.4.1 Transitivity 12
2.2.4.2 Symmetry 12
2.2.4.3 Density 12
2.2.4.4 Diameter 13
2.2.4.5 Structural balance 13
2.2.4.6 Network efficiency 13
2.2.5 Structure highlight algorithms 13
2.2.5.1 Cliques and N-cliques 13
2.2.5.2 N-Clans and N-Clubs 14
2.2.5.3 K-Plexes and K-Cores 14
2.2.5.4 LS-Sets and Lambda Sets 14
3 Military context constraints 17
3.1 Computational complexity tolerance 17
3.2 Topological and flow constrains 19
3.3 Missing data risk 19
4 Algorithms classification 21
5 Conclusion 25
References 27
List of symbols/abbreviations/acronyms/initialisms 29

List of figures

Figure 1 Classification of Analyses of Social Networks 1
Figure 2 Distribution of the analyst questions with respect to the time allowed to provide answers. 18
Figure 3 Functional Flow Diagram depicting the flow of SNA Activities 18
Figure 4 Algorithms classification 25

List of tables

Table 1 Computational complexity scale description 4
Table 2 Topological and flow requirements scale description 5
Table 3 Missing data robustness scale description 6
Table 4 Structure highlight algorithms risk assessment 21
Table 5 Graph measures risk assessment 22
Table 6 Node partitionning algorithms risk assessment 22
Table 7 Centrality and node measures risk assessment 23

1 Introduction

In 2010, DRDC Valcartier engaged in an applied research project called Social Network Analysis in Counter-Insurgency context (SNAC) that aimed at enhancing the intelligence analysis capability. The research project investigated the usage of Social Network Analysis (SNA) techniques, tools and methods in the context of irregular warfare.

While SNA was identified as a must have capacity for the Canadian Forces (CF) and the allies, there is still a lot of confusion on what SNA really is and what it can or cannot do. The capacity sought for by potential users in the field (military, law enforcement, intelligence agencies) is far from what theoretical SNAs can actually provide. In fact, most theoretical SNAs may not, in their actual form, be the single right tool to address the CF needs. The research context where SNA is typically applied is more organised and controlled (close world) than the intelligence one (open world). For this reason, a direct transposition is not possible.

However, this does not mean that theoretical SNAs cannot be put to good use for the intelligence community. If used correctly, they can provide a solid basis for an SNA capability. It is therefore necessary to know the contexts and conditions where SNA is applicable. As mentioned in [1] and [2], failure to do so will provide meaningless answers and could lead intelligence analysts on wrong paths, or provide no answer at all.

To prevent this situation from happening in the SNAC proof-of-concept prototype, efforts were made to clearly identify the limits of application of the analyses performed over social networks in an intelligence context. This document presents risks assessment for common theoretical SNAs (identified in a previous document [3] and depicted as the red set in Fig. 1 as well as the tolerance of the intelligence context to these risks.

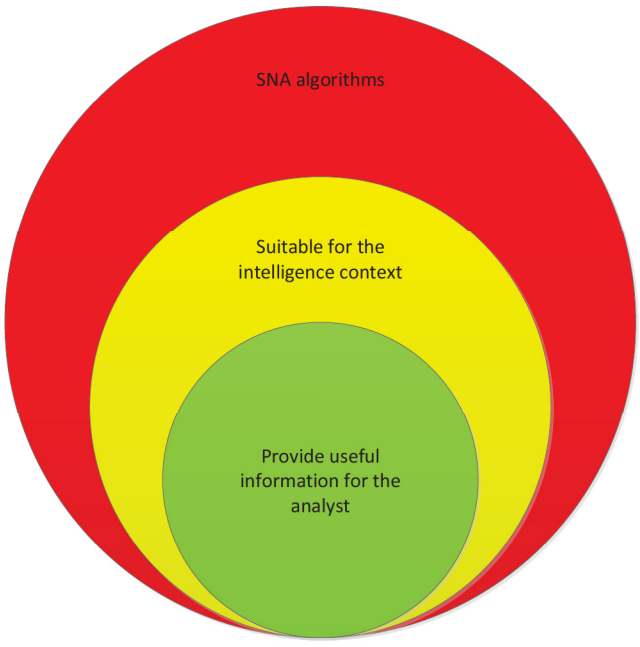

Figure 1 Classification of Analyses of Social Networks

This document presents in chapter 2 the most common SNAs as well as a risk assessment based on three criteria: computational complexity, topological and flow constrains, and missing data robustness. The military intelligence context is also situated using these criteria in chapter 3 to know what it can tolerate in terms of risks. These results will be used in chapter 4 to select the most appropriate SNAs to answer intelligence analysts' questions in chapter 4, i.e., the yellow subset in Figure 1. In chapter 5, recommendations for future research avenues for the SNAC project are provided and include the identification of SNAs that can provide useful information to analysts, i.e., the green subset in Fig. 1.

2 Categorisation of SNA techniques

A lot of research has been done in SNA. Analyses, metrics and algorithms have been found to extract information from the structure of social networks. However, SNAs have preconditions that must be met to be applicable or to provide valid results. In academic work, these constraints pose no problem because work is more theoretical (especially regarding the close world conditions). Unfortunately, in practice, real applications (open world) cannot escape or ignore those constraints. Since this project aims at finding analyses with preconditions that can be met in the intelligence context, this section provides a classification method appropriate to fulfill this objective. A set of evaluation criteria is presented, followed by the assessment of commonly used SNA techniques.

2.1 Evaluation method and criteria

The evaluation method proposed considers three criteria for each SNA. The chosen criteria have been identified in the literature as important aspects to consider before selecting an analysis. They are:

- Computational complexity;
- Topological and flow constraints;
- Missing data robustness.

Combining these criteria provides a relatively good assessment of the difficulty of implementing a particular analysis. Moreover, contexts of application may hold higher risk tolerance for certain criteria. For this reason, the tolerance of the domain of application to risk is also evaluated using the same method. It permits a fast identification of suitable SNA techniques for a particular domain by picking techniques where all the criteria meet the requirement of the domain. The qualitative evaluation scale uses the classes low, medium and high. However, each criterion has its own definition for this scale.

2.1.1 Computational complexity

This criterion refers to the amount of computational resources needed to perform an analysis. It can be related to the algorithm complexity or to the memory needed to complete the analysis. In SNA, this can be function of the number of nodes, the number of links or both. Assuming that the memory is infinite and time is not constrained, no analysis is impossible to compute.

A low risk analysis can be easily computed and presents no real implementation challenges. At medium risk, the analysis may not be suitable in all circumstances, e.g., depending on the size of the graph. High risk analyses should be ignored. The computational complexity scale is explained in Table 1.

Table 1 Computational complexity scale description

Computational complexity	Characteristics
Low	• Algorithm complexity is below or equal to O(n^2) • Suitable for all sizes of graphs • Can be used for human interaction
Medium	• Algorithm complexity greater than O(n^2) • Can be used for human interaction for small graphs. • Computation on large graphs must be made offline
High	• Algorithm complexity exponential or O(n^x) with large x. • Computation must be made offline for small graphs. • Almost impossible to use for large graphs. • Number of results too large to fit in memory.

2.1.2 Topological and Flow requirement

Social networks can have many different topologies, also called configurations in this document. They also represent the connections between entities using different types of relations. In some cases, these relations characterise the flow inside the network. This evaluation criterion measures if an analysis is versatile enough to provide valuable information on a network no matter what the topology or the type of flow of the network is.

A low risk analysis can be used on any topologies or flows. At medium risk, analyses may not be suitable in all circumstances. They may necessitate a special topology or they only provide correct results for certain types of flow. High risk analyses are highly specialised and are too restrictive to be used easily in different contexts. The return on investment may not be there and they should potentially be ignored. The topological and flow requirements risk scale is explained in Table 2.

Table 2 Topological and flow requirements scale description

Topological and flow requirements	Characteristics
Low	• Algorithm can handle all types of network configurations • Algorithm provides useful information on nearly all types of flows.
Medium	• Algorithm depends on often seen network configuration. • Algorithm only provides useful information for certain types of network flows.
High	• Algorithm depends on not often seen network configurations. • Algorithm depends on often seen network configuration and only provides useful information for certain types of network flow.

2.1.3 Missing data robustness

Usually in the literature reviewed, analysis of SNs is typically performed on networks devoid of missing data. This is an implicit assumption in most academic work and conclusions provided through analyses. Unfortunately, this assumption does not always hold since social networks are hard to observe, specifically for covert networks. This criterion is important because it helps qualify the level of resilience of analyses.

A low risk analysis is barely affected by missing data. At medium risk, an analysis may not be suitable in all circumstances. It may withstand some missing data or the results may be a bit off. High risk analyses should be ignored in some specific cases, or else provided with significant caveats. The missing data robustness scale is explained in Table 3.

Table 3 Missing data robustness scale description

Missing data robustness	Characteristics
Low	• Algorithm can handle all types of missing data • Algorithm provides sound results even with a lot of missing data.
Medium	• Algorithm can handle missing links or nodes • Algorithm still provides sound results with small amount of missing data. • Algorithm provides sound results with missing data, but the interpretation may be different.
High	• Algorithm provides meaningless results in presence of missing data.

2.2 SNA algorithm assessment

This section provides the risk assessment, along with justifications, for analyses commonly performed on social networks. Selected analyses and their brief descriptions come from previous work on the SNAC project. Since they are described in details in a separate document [3], this section does not give complete descriptions and interpretations of them. It is important to note that the classification presented here is subjective since there is no mean to quantify the risks. This assessment is intended to rank analyses in order to lower the risk of implementing them in a specific context where risk tolerance is low. A higher risk does not mean that an analysis could not be useful; it means that developing it may require a deeper evaluation of the return on investment. For example, a method may have a high risk because it is too specific to be used in other context than the one it was meant for initially.

2.2.1 Centrality measures

Centrality measures aim at identifying important nodes in a network. They rank nodes with respect to how central they are, based on selected characteristics like their number of connections, their distance from important nodes, and so on. Since they are versatile in term of assessment, they have been one of the most used tools in applied SNA. For this reason, some of their limitations are well known and documented. The risk assessment provided below synthesises some of these documented issues.

2.2.1.1 Degree centrality

Degree centrality, also called prestige or in-degree and out-degree for oriented graphs, is the simplest measure of all. It just counts the number of edges adjacent to a node. The complexity of this measure is low since it is linear. Moreover, there are no topological and flow restrictions. However, it is sensitive to missing links. If the missing data about links is uniform over the network, then it is not much of a problem. In all other cases, the result may be biased to an unknown level. For these reasons, the risk assessment for degree centrality for each of the three criteria is: low, low, and low.

2.2.1.2 Eigenvector centrality

The centrality score associated with a node is the sum of the centrality scores of its connected nodes. An eigenvector approximation (since one cannot compute the exact value) is used to solve this recursive relation. Eigenvectors of a matrix represent vectors that do not change orientation when multiplied to the said matrix. They can be used to represent the balance point of the mutual influence between nodes, i.e., to solve the recursive relation. PageRank and Bonacich Power, Friedkin total effects, and Katz influence are variations on this measure and, as eigenvector centrality, can be used in signed and valued graphs. The complexity of this measure is medium since it is solving a simple recursive relation that can take a lot of iterations to converge. In addition, there are no topological and flow restrictions. It is sensitive to missing links as the recursive nature of this measure propagates the impact of a single missing link over the network. Also, a single missing link between nodes that themselves differ greatly in importance can have a major effect. For these reasons, the risk assessment for Eigenvector centrality for each of the three criteria is: medium, low, and medium.

2.2.1.3 Freeman closeness

Closeness centrality of a node is the sum of the length of the geodesics to every other nodes. The complexity of this measure is medium since it is between O(n^2) and O(n^3), depending on the density of the graph. It can only be applied to connected networks and the flow is restricted to the shortest path. It is sensitive to missing links as their impact of can spread over large parts of the network. Also, a single missing link between nodes can hide an important shortest path and have a major effect on the measure. For these reasons, the risk assessment for closeness centrality is, for each of the three criteria: medium, high, and medium.

2.2.1.4 Topological centrality

This measure is an iterative process like eigenvector centrality. Nodes and edges are attributed with initial weights and the iterative process adjusts them based on the surrounding components. So centrality scores that are given to nodes and edges are not only function of the topology of the graph, but also of the initial weights. The complexity of this measure is medium since it is solving a simple recursive relation that can take a lot of iterations to converge. There are no topological and flow restrictions. It is sensitive to missing links as the recursive nature of this measure propagates the impact of a single missing link over the network. Also, a single missing link between nodes that themselves differ greatly in importance can have a major effect on the measure. The same problem can be observed if initial weights for nodes and links are not carefully chosen since they quantify their importance. For these reasons, the risk assessment for topological centrality is, for each of the three criteria: medium, low, and medium.

2.2.1.5 Freeman Betweenness

Betweenness sums up the number of geodesics a node is part of when solving the all-sources shortest path problem. In the case of multiple geodesics between two nodes, the contribution is divided evenly. The complexity of this measure is medium since it is between O(n^2) and O(n^3), depending on the density of the graph. It is only suitable for flows that use the shortest path. It is sensitive to missing links as their impact can spread over large part of the network. Also, a single missing link between nodes can hide an important shortest path and have a major effect on the measure. For these reasons, the risk assessment for freeman betweenness centrality is, for each of the three criteria: medium, medium, and medium.

2.2.1.6 Bridging centrality

A bridging node is a low degree node used by many geodesics, but that is surrounded by high degree neighbours. So the bridging centrality of a node is the product of the betweenness centrality and a bridging coefficient that has high value when a node is linked to higher degree nodes, and low values when linked to lower degree nodes. The complexity of this measure is medium since it is between O(n^2) and O(n^3), depending on the density of the graph. It is only suitable for flows that use the shortest path. It is sensitive to missing links as their impact can spread over large parts of the network. Also, a single missing link between nodes can hide an important shortest path and have a major effect on the measure. For these reasons, the risk assessment for bridging centrality is, for each of the three criteria: medium, medium, and medium.

2.2.1.7 Flow Betweenness

Given that a node inside a directed graph has a capacity (i.e., the flow that can transit trough it), flow betweenness of a node is defined as the average flow obtained when solving the maximum flow problem between all possible sources and sinks. The complexity of this measure is medium since it is between O(n^3) and O(n^4), depending on the density of the graph. There is a required capacity for all nodes, but uniform values are suitable. It is sensitive to missing links, but less than betweenness and closeness. The impact of missing links can spread over large parts of the network. For these reasons, the risk assessment for flow betweenness centrality is, for each of the three criteria: medium, low, and medium.

2.2.1.8 Dependence centrality

Dependence centrality represents how many times a node is used when computing a single source shortest path. The result is a matrix where values sum up the number of times a node A was used by geodesics from B to reach all other nodes in the network. The sum of each row of the matrix gives the same result as betweenness centrality. The complexity of this measure is medium since it is between O(n^2) and O(n^3), depending on the density of the graph. It is only suitable for flows that use the shortest path. It is sensitive to missing links as their impact can spread over

large parts of the network. Also, a single missing link between nodes can put out of sight an important shortest path and have a major effect on the measure. For these reasons, the risk assessment for dependence centrality is, for each of the three criteria: medium, medium, and medium.

2.2.1.9 Position role centrality

This measure depends on the basic definition of efficiency, which is the harmonic mean of the shortest distance between each pair of nodes. This centrality measure of a node is the difference between the efficiency of the network with and without this node. The complexity of this measure is medium since it is between O(n^3) and O(n^4), depending on the density of the graph. It is only suitable for flows that use the shortest path. It is sensitive to missing links as their impact can spread over large parts of the network. Also, a single missing link between nodes can put out of sight an important shortest path and have a major effect on the measure. For these reasons, the risk assessment for position role centrality is, for each of the three criteria: medium, medium, and medium.

2.2.1.10 Newman Random walk centrality

Random walk centrality quantifies how central a node is regarding the opportunity to receive information that would have been randomly diffused over the network. The complexity of this measure is medium since it is between O(n^3) and O(n^4), depending on the density of the graph. It is sensitive to missing links but less than betweenness and closeness centrality measures since it uses all possible paths and not only geodesics. The impact of missing links can spread over large parts of the network. For these reasons, the risk assessment for random walk centrality is, for each of the three criteria: medium, low, and medium.

2.2.2 Other node measures

Beside centrality there are other measures that quantify different aspects of a node within a network. Equivalence between nodes is one aspect, along with the roles and the density of ego networks for instance. Selected measures are evaluated below.

2.2.2.1 Structural equivalence

Two nodes are structurally equivalent if they are linked to the exact same nodes in the same way. However measures of equivalence are preferred since they provide a degree of equivalence instead of a Boolean value in the exact case. There exists two main approaches commonly used to indexing pairwise structural equivalence in order to highlight similarities. They are the correlation coefficient and the squared Euclidean distance. The complexity of these measures is low since finding equivalents for a node is a linear procedure O(n) and, finding all pair equivalents is thus O(n^2). The impact of missing links is low; it does not spread over large parts of the network and the impact is mostly local. For these reasons, the risk assessment for structural equivalence is, for each of the three criteria: low, low, and low.

2.2.2.2 Automorphic and Isomorphic equivalence

Automorphic equivalent nodes occupy positions that have exactly the same perspective on the structure of the network, i.e., there is no way to differentiate them in terms of positions in the network. Isomorphic equivalence is the same concept, except that nodes reside on different graphs. There is no known fast algorithm that guarantees the identification of equivalent nodes in all graphs. However, since equivalent nodes occupy indistinguishable structural locations in a graph, one can suggest potential equivalent candidates by looking at graph theoretic measures. Selecting pairs of nodes on the basis that they have the same centrality on many centrality measures is a good way to approximate this equivalence. The complexity of the computation therefore depends on the graph theoretic measures being selected. Also, for automorphic nodes (but not for isomorphic equivalent nodes) must be inside the same cliques and subgroups. The complexity of isomorphic equivalence relies on a combination of centrality measures, so the computational complexity may be evaluated to be around O(n^5). Since one has to find cliques to determine automorphic equivalence; the complexity is therefore exponential. There is no topological or flow restriction for these measures since they look at the topology itself to tell nodes apart. These measures are sensitive to missing links as they change the topology of the network. For these reasons, the risk assessment for isomorphic equivalence is, for each of the three criteria medium, low, and medium, while for automorphic equivalence it is and high, low, and medium.

2.2.2.3 Regular equivalence

With respect to this type of analysis, two nodes are said to be equivalent if they are linked to the same classes (or roles) of nodes. The number of ties is irreverent. Equivalent nodes just need to have at least one link to the same classes. Finding subsets of regularly equivalent nodes is a recursive analysis where the role of a node is defined by the roles of its relations. The REGE algorithm exists to compute this measure and uses an iterative procedure within which estimates of the degree of regular equivalence between pairs of actors are adjusted in light of the equivalences of the adjacent nodes. The complexity of this algorithm is medium O(n^3). It also requires valued oriented graphs to provide non trivial answers. It is sensitive to missing links and their impact can spread over large parts of the network due to the recursive nature of the procedure. For these reasons, the risk assessment for regular equivalence is, for each of the three criteria: medium, medium, and medium.

2.2.2.4 Local role equivalence

Local role equivalence is based on the concept that similar nodes are related to other nodes using the same combination of relations in a multigraph, i.e., the set of possible classes of paths between two nodes is the same. This measure is conceptual and no real application of it is documented, possibly because of the complexity of the problems. For instance, it is costly to construct an enumeration of all possible paths between nodes. This task is done by costly matrix multiplications and is highly affected by missing values. For these reasons, the risk assessment for local role equivalence is, for each of the three criteria: high, medium, and high.

2.2.2.5 Brokerage

Given a graph and a partition of nodes, this algorithm quantifies how close a node is to being one of five kinds of brokerage types: coordinator, consultant, gatekeeper, representative or liaison. These classes represent the number of possible combinations of group membership of three nodes involved in a non-transitive relation. This quantification of relation is pretty useful to evaluate the roles of a node in a graph containing many factions. The time complexity of this method is O(n^3). This method also needs a partitioned network where nodes can only be associated with one group. It is robust to missing data because it is computed using triads so missing links have small local effects. For these reasons, the risk assessment for brokerage is, for each of the three criteria: medium, medium, and low.

2.2.2.6 Clustering coefficient

The clustering coefficient measures the proportion of triads a node is part of in regards to all the possible triads its ego network can generate. When this value is close to one, this measure means that the ego network is close to being a clique. A value closer to zero means that the node is a structural hub (star network) inside this ego network. The complexity of this measure is O(n^2) and there is no topological requirement. The impact of missing links is low as it does not spread over large parts of the network and the impact is local to the ego network of the connected node. For these reasons, the risk assessment for clustering coefficient is, for each of the three criteria: low, low, and low.

2.2.3 Node partitioning algorithm

This section describes methods for partitioning nodes into subsets so that nodes within each subset are closer to being equivalent than nodes in different subsets are. The utility of partitioning is when it is nearly impossible to see any pattern in a similarity matrix. Clustering nodes together helps one understand the informal organisation in the graph.

2.2.3.1 CONCOR

CONCOR means convergence of iterative correlation. It is a means of splitting a socio-matrix into two subsets by successively calculating the correlation between its rows until all values converge to 1 or -1. The resulting subsets contain nodes with stronger correlation than nodes taken from different subsets. The process can be reused on subsets to create a hierarchy. The complexity of the algorithm is O(n^3). Repeated use of CONCOR to split a graph cannot be worse than O(n^4). There is no topological requirement. However, missing links can lead to local misclassification. For these reasons, the risk assessment for CONCOR is, for each of the three criteria: medium, low, and medium.

2.2.3.2 Hierarchical clustering

Hierarchical clustering is an iterative process to cluster two-by-two graph subsets. The procedure starts with subsets containing only one node and merges the most similar ones. As opposed to CONCOR, instead of splitting nodes, hierarchical clustering groups them. The algorithm is performed on a matrix of similarity between nodes. Any similarity measure can be used. The

complexity of the algorithm is O(n^3), and there is no topological requirement. Missing links are not an issue since similarity values are provided. For these reasons, the risk assessment for hierarchical clustering is, for each of the three criteria: medium, low, and low.

2.2.3.3 Blockmodels

A blockmodel clusters together nodes that are mostly structurally equivalent and infers potential relations between these clusters. The algorithm tries to find the optimal solution based on a selected similarity criterion by permuting positions inside the socio-matrix. A blockmodel can be interpreted as a high-level graph of relations between roles. As stated in the Pajek software [4], the real limit of the blockmodel algorithm is the time complexity. On 100 vertices, the procedure can last some hours. The exact complexity is not given. There is no topological requirement. However, missing links can induce local misclassification. For these reasons, the risk assessment for blockmodeling is, for each of the three criteria: high, low, and medium.

2.2.4 Graph measures

Graph measures quantify global characteristics of networks. They are not useful in describing a particular node, but can provide some insight about the environment where it is evolving. They can also be used to monitor communities over time for change detection.

2.2.4.1 Transitivity

Transitivity is defined as “if A is related to B and B is related to C, then A is related to C”. Network transitivity is the proportion of actual transitive relations relative to the total number of possible ones. The complexity of the algorithm is O(n^2). There is no topological requirement. The quantity of missing links is directly proportional to the potential error of the measure. Hence, some missing links are not problematic but can slightly change the overall result. For these reasons, the risk assessment for transitivity is, for each of the three criteria: low, low, and low.

2.2.4.2 Symmetry

Symmetry is defined as “if A is related to B then B is related to A”. Network symmetry is the proportion of actual symmetric relations relative to the total number of possible ones. The complexity of the algorithm is O(n). This metric only makes sense for oriented graph since non-oriented graphs are always symmetric. The quantity of missing links is directly proportional to the potential error of the measure. Hence, some missing links are not problematic but can slightly change the answer. For these reasons, the risk assessment for symmetry is, for each of the three criteria: low, medium, and low.

2.2.4.3 Density

Network density is the proportion of relations in a network relative to the total number of possible relations. The complexity of the algorithm can vary from O(1) to O(n^2) depending on how relations are physically stored (lists or matrices). There is no topological requirement. The quantity of missing links is directly proportional to the potential error of the measure. Hence,

some missing links are not problematic but can slightly change the answer. For these reasons, the risk assessment for density is, for each of the three criteria: low, low, and low.

2.2.4.4 Diameter

The diameter of a connected graph is equal to the length of its longest geodesic between any pair of nodes. The diameter analysis relies on the algorithm to find all sources shortest path which is at most O(n^3). The network needs to be connected since the all sources shortest path cannot be used on disconnected network. Missing links are not an issue in most cases but can be problematic on networks with low density as this would increase the chance of creating shortcuts. For these reasons, the risk assessment for diameter is, for each of the three criteria: medium, high, and low.

2.2.4.5 Structural balance

Given a network where relations represent positive and negative relation, it is possible to check if the network is coherent. To do so, the structural balance algorithm verifies if all possible cycles in the graph have the same polarity. Balanced networks are considered more stable, while unbalanced networks are candidates for change through seeking balance. To compute this measure one must perform successive matrix multiplications until non-balance is found, or until the matrix has been elevated to the power of the length of the longest cycles. Thus, the complexity of this measure is polynomial and, in the worst case, a balanced large graph can be costly to compute. The network also needs to be connected by positive and negative relations. Finally, a single missing value can change the balance of the network. For these reasons, the risk assessment for structural balance is, for each of the three criteria: high, medium, and high.

2.2.4.6 Network efficiency

The network efficiency is a measure to quantify how efficiently the nodes of a network exchange information. It is the harmonic mean of the geodesic distance between all pairs of nodes. The complexity of this measure is medium since it is between O(n^2) and O(n^3). There is no topological requirement. A single missing link between nodes can hide an important shortest path and have a major effect on the measure. For these reasons, the risk assessment for network efficiency is, for each of the three criteria: medium, low, and medium.

2.2.5 Structure highlight algorithms

The algorithms evaluated in this section try to identify dense subsets of the network. Based on selected criteria, the nodes in these subsets are close to one another, as for instance the number of connections or the distance from other nodes within the subset.

2.2.5.1 Cliques and N-cliques

A clique is a subset of individuals in which every person is connected to every other persons. While this concept may be of interest, it is highly prone to missing data errors. An N-clique is a more relaxed algorithm where members of the group need only to be connected by geodesics of maximum length N to other members.

Also, the sub-graph needs to be maximal. This less strict version is more suitable for graphs with potentially missing data. Surprisingly, in an n-clique, members can be connected through non-members. Indeed, a node can be in many cliques or n-cliques at the same time. The complexity of the algorithms that highlight these sets is exponential, making them unusable for most graphs. There is also no topological requirement for either problem. For these reasons, the risk assessment for clique finding is, for each of the three criteria: high, low, and high, while it is high, low, and medium for N-clique identifying.

2.2.5.2 N-Clans and N-Clubs

A kind of sub-graph similar to that of the n-clique is that of n-clans and n-clubs. The intention here is to improve the cohesion of n-cliques by introducing some restrictions. For n-clans, the geodesic distance in the n-cliques must reside within the sub-graph. No members can be connected by non-members. An N-club is a maximal sub-graph of diameter N. N-clubs are not necessarily n-cliques, but are always sub-graphs of n-cliques. A node can be in many n-clans or n-clubs at the same time. The complexity of the algorithms that highlight these sets is exponential, making them unusable for most graphs. There is also no topological requirement for either problem. Missing links can change the size and the membership of the sub-graph. For these reasons, the risk assessment for N-clans and N-clubs is, for each of the three criteria: high, low, and medium.

2.2.5.3 K-Plexes and K-Cores

Cliques, clans and clubs are based on the concept of complete members' connections, itself based on the relative geodesic distance. Another way to look at cohesion is to use as a criterion the number of ties within the sub-graph. N-cliques, n-clans and n-clans are flexible versions of cliques in term of distance. K-plexes and k-cores are flexible versions of cliques in terms of connection. K-plexes are maximal sub-graphs containing N nodes in which each node is adjacent to no fewer than N-K nodes in the sub-graph. A k-core is a sub-graph in which each node is adjacent to at least a minimum number of k other nodes in the sub-graph. A node can be in many k-plexes or k-cores at the same time. The complexity of the algorithms that highlight these sets is O(n^3) for K-cores and exponential for K-plexes. There is also no topological requirement for either problem. Few missing data can slightly change the size and the membership of the sub-graph. For these reasons, the risk assessment for K-cores is for each of the three criteria: medium, low, and medium, while it is high, low, and high for K-plexes.

2.2.5.4 LS-Sets and Lambda Sets

These two types of subsets are based on the comparison of ties within and outside the set. This type of measure is based on the observation that cohesive groups are more closely tied to each other than to outsiders. A sub-graph is an ls-set if each of its subsets has more ties to its complement within the sub-graph than outside of it. In a lambda set, any pair of nodes in the set has greater line connectivity than any pair of nodes consisting of one node in the set and one outside of it. The line connectivity of two nodes is defined as the number of paths that contain no relation in common. By definition, Ls-sets and Lambda sets do not overlap. The complexity of the algorithms that highlight these sets is O(n^4), making them unusable for most graphs.

There is also no topological requirement for either algorithm. A few missing data can slightly change the size and the membership of the sub-graph. For these reasons, the risk assessment for LS-Sets and Lambda Sets is, for each of the three criteria: medium, low, and medium.

This page intentionally left blank

3 Military context constraints

The military intelligence context has its own constraints that a SNA capability must take into consideration. This section provides an assessment of these constraints using the criteria provided earlier. The assessment is based on knowledge of military procedures acquired during meetings, workshops and other elicitation activities done for the SNAC project. This elicited knowledge of military intelligence requirements could obviously benefit from additional information. Nevertheless, it establishes a legitimate starting point to provide a solid lower bound estimate for risk assessment. This assessment will be used to sift through the analyses of SNs and retain those that are more suitable to the military context.

3.1 Computational complexity tolerance

To be efficient, the military intelligence process needs information within a certain timeframe, depending on the activities being considered. An SNA capability may provide answers to critical questions. However, the algorithms to generate them can, in some instances, be too complex to provide answers within the timeframes. The challenge here is to ascertain ways to determine what is the maximum complexity an algorithm is allowed to reach while still being useful. This question is difficult to tackle because it depends on many factors like the size of the dataset, the processing power available, the time available to provide an answer, and the value of the information. Using knowledge of the military context as captured in a previous document [5], it is still possible to set a lower bound on the maximum allowable complexity of SNA algorithms.

The first fact is that answers for military intelligence must be provided in a specified timeframe. The above mentioned report [5] lists nearly 500 questions of interest for the analyst in a COIN context, with their required timeframe. Figure 2 summarizes the proportion of questions for each timeframe. It is clear that most of the questions of interest, 53%, need to be answered in less than a day. The second fact is that the analyst must go through a specific process to provide answers (see Fig 3). Based on this, the overall timeframe must include the time to understand the question, find the data, find the answer, provide justification and format the document. The time to run SNA algorithms must fit inside this functional work flow. So if the timeframe is for example 5 hours, the algorithm may have in reality less than an hour to provide its result. The third and final element is that highly complex algorithms require dedicated resources. It may be possible to run thousands of sub-exponential SNA algorithms in the same time window required to perform a single exponential SNA.

For these reasons, the military intelligence process should limit itself to the use of algorithms with low and medium computational complexity risks.

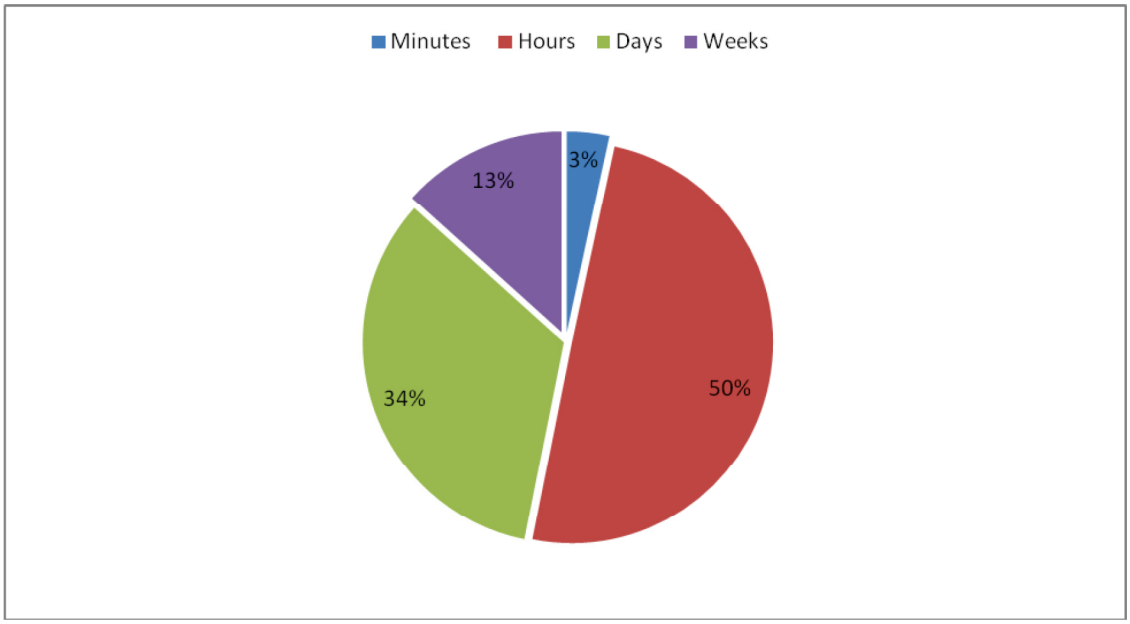

Figure 2 Distribution of the analyst questions with respect to the time allowed to provide answers.

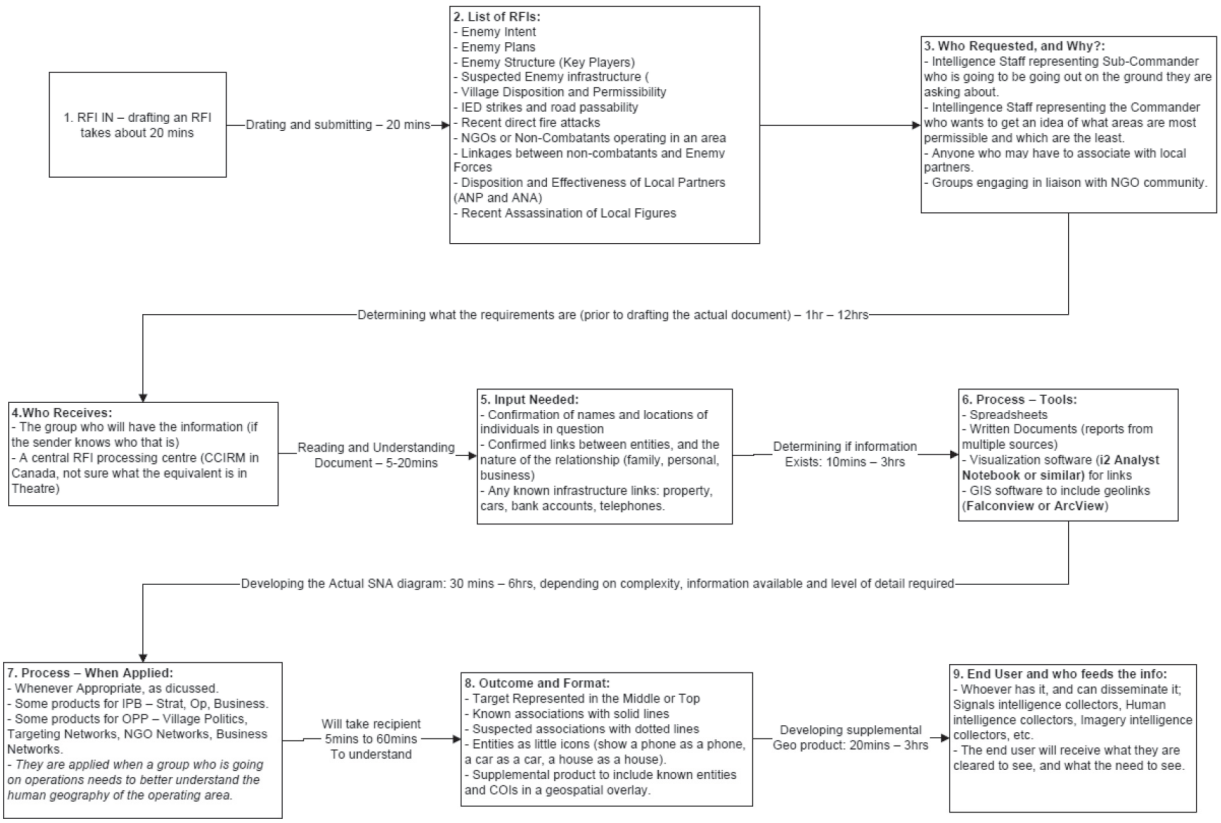

Figure 3 Functional Flow Diagram depicting the flow of SNA Activities

3.2 Topological and flow constrains

Some analyses of SNs need special network configurations or representations in order to be used. Some need a network with a very specific network flow for the analyses to provide meaningful answers. Both situations put constraints on the applicability of SNA by imposing specific topologies or semantics on the network. Arguments to determine the lower bound on tolerance of the intelligence process regarding this matter are closely related to the medium used to store and disseminate information, e.g. using unstructured text documents.

Unstructured data sources are difficult to exploit as their lack of structure makes any automated data extraction and processing difficult. However, such a lack of structure is compensated by rich expressivity and flexibility. They can describe with rich semantics the nature of the relations inside social networks. By reading a text report, an analyst can more easily determine if a network is oriented, bi-modal, valued, etc., so the information regarding the flow is well captured. However, some topological constraints like network connectedness cannot be addressed in any way since they depend on the network itself.

Since network flow information is well captured by military intelligence, social network flow constraints do not pose a problem. However, a number of topological configurations required by some SNAs are data dependent and cannot be met. Since the networks handled by the analysts can take any form, it is impossible to always provide a suitable topology for an analysis. For these reasons, military intelligence process should limit itself to the use of low and medium topological and flow risk algorithms.

3.3 Missing data risk

An objective of the intelligence process is to provide information about the current situation being faced. Since a continuous process can be considered as a sequence of steps repeated over time, the intelligence process provides a continuous flow of information discovered over time. However, this process is always in a catch-up mode, and the information flow coming from the different sources has its limitations.

The intelligence process may first build an initial awareness of the environment. In this step, the learning curve is steep and the information flow is high; the understanding of the social network evolves rapidly and it may appear to grow and morph as missing data is reduced. In this initial phase, it is believed that even robust SNAs can be misleading. The assumption is that when the knowledge base of the network reaches a certain size, the information flow specific to the network under study will reduce. At that point, new information relates more to changes in the network. In this phase, most SNAs may be applicable and should provide more constant and reliable results. However, this second situation assumes that the social network representation is a reliable representation of the situation. There may remain covert actors or relations that have the potential to change results of SNAs. Unfortunately, there is no reliable way to assess the goodness of fit of the network representation and so, the level of missing data will almost always be unknown.

This last consideration makes the use of SNA in the military context a bit of a gamble. One may never know if the answer provided by an algorithm may be wrong. To provide results that the

military analysts may rely upon, SNAs should be robust to missing data. For these reasons, military intelligence process should limit itself to the use of low missing data risk algorithms.

4 Algorithms classification

This section summarizes the risk evaluation of the algorithms and the lower bounds on the tolerance of the intelligence process to these risks. The following tables synthesise the results of the assessments of the previous sections. Each risk is color-coded in accordance with in the description of the risk previously provided. Also, the colors used for the name of the algorithm represent whether this algorithm is suitable for the military context. Green represents algorithms with risks below or equal to the tolerance of the intelligence process. They can be used without any particular consideration. Yellow is for algorithms with risks above the tolerance, but without any high risk assessment for any of the criteria. Therefore, caution is advised when planning to use these algorithms if one wants usable results. They may take too long to execute, have fancy data requirements or provide meaningless results in presence of missing data. Red is for algorithms with at least one high risk assessment and that are not suitable for real life applications.

Table 4 Structure highlight algorithms risk assessment

Method	Computational complexity	Topological and flow requirement	Missing data robustness
Cliques			
N-cliques			
N-Clans			
N-Clubs			
K-plexes			
K-cores			
LS-Sets			
Lambda sets			

Table 5 Graph measures risk assessment

Method	Computational complexity	Topological and flow requirement	Missing data robustness
Transitivity			
Symmetry			
Density			
Diameter			
Structural balance			
Network efficiency			

Table 6 Node partitionning algorithms risk assessment

Method	Computational complexity	Topological and flow requirement	Missing data robustness
CONCOR			
Hierarchical clustering			
Blockmodeling			

Table 7 Centrality and node measures risk assessment

Method	Computational complexity	Topological and flow requirement	Missing data robustness
Degree centrality			
Eigenvector centrality			
Freeman closeness			
Topological centrality			
Freeman Betweenness			
Bridging centrality			
Flow Betweenness			
Dependence centrality			
Position role centrality			
Newman Random walk centrality			
Structural equivalence			
Isomorphic equivalence			
Automorphic equivalence			
Regular equivalence			
Local role equivalence			
Brokerage			
Clustering coefficient			

This page intentionally left blank

5 Conclusion

This document provided an assessment of the risk of using some of the well-known SNA analyses based on three criteria: computational complexity, topological and flow requirements and missing data robustness. The tolerance of the intelligence process to these risks was also evaluated. Using the risk evaluation of the algorithms and the lower bound on the tolerance of the intelligence process to these risks, it was possible to identify 8 suitable algorithms, 11 unusable and 15 risky ones out of 34 algorithms (see Fig. 4). In light of these results, it is possible to draw some conclusions to orient future work.

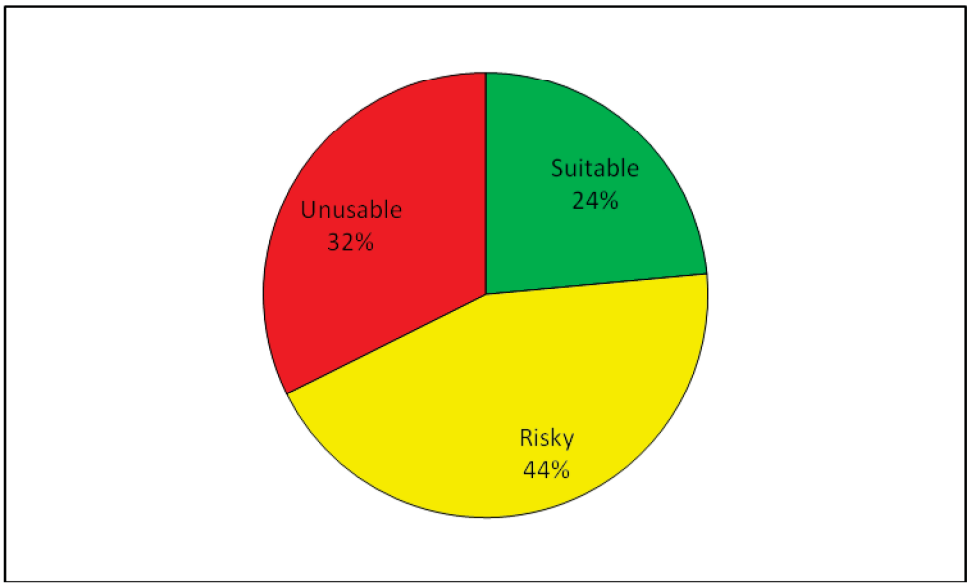

Figure 4 Algorithms classification

First, existing and commonly used analyses of SNs are not sufficient to fulfill the needs of the intelligence function. Some of these algorithms can be of use, but further investigation is required to identify if, in the envisioned context, they can provide useful results. Indeed, a second examination of them is needed to ensure they provide useful information for the analyst match, i.e., they belong to the green subset of Fig. 1. The remaining algorithms appear to be too complex or too sensitive to the missing data issue.

Second, missing data is a major problem. Most of SNAs cannot provide correct results under missing data conditions. However, since the source and gravity of missing data can differ greatly, maybe some SNAs can still be used under certain circumstances. Further study on this topic should look at types of missing data and confidence interval for the given results. A potential avenue to circumvent this problem is to restrict analyses to small complete sub-graphs, i.e., triad or quad. In this case, missing data are by definition absent and computational complexity is not an issue. This scheme would make almost all analyses usable (some with a different interpretation), except for structure highlight algorithms, which are then not needed due to the small size of the graph.

Third, grouping in a social network is a complex task for a computer. Most of the known algorithms take exponential time to solve this problem (see Table 4), and even approximations are also proven to be exponential. However, maybe there is no need for a perfect answer when it comes to finding sub-graphs of interest. The analyst may also provide assistance for these tasks. Technologies like visual analytics should be investigated to solve parts of this problem.

Finally, social networks contain a lot of information that could be useful for the analyst. However, analysts have a lot of questions that need answers. Useful information provided by the eight (8) suitable algorithms identified in this document is not enough to fill this gap. Future work should focus on developing *ad hoc* analyses to answer military needs. Providing complete, reliable and timely information is one goal of the SNAC project that theoretical SNA cannot fulfill alone.

References

[1] Wasserman, S., and Faust, K., (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press. ISBN 0-521-38269-6.

[2] Borgatti, S. P., (2005), *Centrality and network flow*, Social Networks, Vol. 27, # 1, pp. 55-71.

[3] Martineau, E. and Lecocq R., *Review of measures for Social Network analysis in COIN context*, UNCLASSIFIED, DRDC Valcartier Technical memo to appear.

[4] V. Batagelj and A. Mrvar, *Pajek – Program for Large Network Analysis*. Home page: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[5] Banbury S. and Miller J., *Study of SNA, IPB/IPOE and COIN SNA Triplet Inventory Report*, UNCLASSIFIED, DRDC Valcartier contract report.

This page intentionally left blank.

List of symbols/abbreviations/acronyms/initialisms

[Enter list here, if applicable. If not, delete the page.]

DND	Department of National Defence
DRDC	Defence Research & Development Canada
DRDKIM	Director Research and Development Knowledge and Information Management
R&D	Research & Development
CF	Canadian Forces
SNA	Social Network Analysis
SNAC	Social Network Analysis in Counter-Insurgency context
COIN	COunter-INsurgency

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.)

Defence R&D Canada – Valcartier
2459 Pie-XI Blvd North
Quebec (Quebec)
G3J 1X5 Canada

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

UNCLASSIFIED
(NON-CONTROLLED GOODS)
DMC A
REVIEW: GCEC April 2011

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

Applicability of Social Network Analyses in an Intelligence Context:

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

Martineau E.

5. DATE OF PUBLICATION (Month and year of publication of document.)

December 2012

6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)

46

6b. NO. OF REFS (Total cited in document.)

5

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.)

Defence R&D Canada – Valcartier
2459 Pie-XI Blvd North
Quebec (Quebec)
G3J 1X5 Canada

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.)

15do

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

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.)

DRDC Valcartier TM 2013-418

10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

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

Unlimited

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.))

Unlimited

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.)

This document presents a classification of theoretical social network analyses to determine if they are suitable for the military intelligence context. This work was done under the project 15do, "Social network analysis in a counter insurgency context", to orient the development of a proof of context prototype. Thirty-four (34) theoretical social network analyses were selected and classified based on the implementation risk they pose. The evaluation criteria used are computational complexity, topological and flow requirements and missing data robustness. The risk tolerance of the military intelligence process was also evaluated using the same criteria. Using both results, it was possible to identify 8 suitable algorithms, 11 unusable and 15 risky ones. These results suggest that classical social network analysis methods are not sufficient to answer the military requirements stated in this project. Alternative ways to extract information and perform network analysis should be investigated. In particular, there is a need for algorithms that are robust to missing data and can extract sub-networks of interest efficiently. The impact of missing data is still not well understood and should be the focus of further research as it is largely encountered in the context of covert networks.

Ce document présente une classification des analyses des réseaux sociaux pour déterminer si elles sont utilisables dans le contexte du renseignement militaire. Ce travail a été effectué sous le projet 15do, « Social network analysis in a counter insurgency context », pour orienter le développement d'une preuve de concept logiciel. Trente-quatre (34) analyses de réseaux sociaux théoriques ont été sélectionnées et classées, basé sur le risque d'implémentation qu'elles posent. Les critères d'évaluation utilisés sont : la complexité algorithmique, les requis topologiques et de flux et la robustesse aux valeurs manquantes. La tolérance au risque du processus du renseignement militaire a été aussi évaluée à l'aide des mêmes critères. En utilisant ces deux résultats, il est possible d'identifier 8 algorithmes utilisables, 11 non-utilisables et 15 à haut risque. Ces résultats suggèrent que les analyses de réseaux sociaux classiques ne suffisent pas à répondre aux besoins des militaires énoncés dans ce projet. Des méthodes alternatives d'extraction d'information et d'analyse de réseaux devraient être recherchées. En particulier, il existe un besoin pour des algorithmes robustes aux données manquantes et qui peuvent extraire des sous-réseaux d'intérêt efficacement. De plus, l'impact des données manquantes n'est pas bien compris et devrait faire l'objet de recherches plus poussées puisque c'est un phénomène souvent rencontré dans le contexte des réseaux secrets.

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.)

SNA

Defence R&D Canada

Canada's Leader in Defence
and National Security
Science and Technology

R & D pour la défense Canada

Chef de file au Canada en matière
de science et de technologie pour
la défense et la sécurité nationale

www.drdc-rddc.gc.ca