Abstract

Data analysis, management, and governance are critical elements of insider threat (InT) programs. To strengthen the InT community's ability to improve their prevention and response efforts, the Threat Lab developed the *Data Analysis Plan for the Insider Threat Community: A Collaborative Workbook* (henceforth, "Workbook"), an instructional guide for use by InT analysts. The research team conducted stakeholder meetings and semi-structured interviews to identify opportunities to enhance data analysis efforts across InT programs. This resulted in an operationally relevant workbook with supplemental material to enable end-users to conduct analyses and interpret results. The results from this effort also highlight the need for policy to standardize capabilities across the Federal Government.

the ThreatLab
Integrating Research Into Practice

About The Threat Lab

The Defense Personnel and Security Research Center (PERSEREC) founded The Threat Lab in 2018 to incorporate the social and behavioral sciences into the insider threat mission space. Our team is headquartered in Seaside, California, and includes psychologists, sociologists, policy analysts, computer scientists, and other subject matter experts committed to workforce protection.

Developing a Data Analysis Plan for the Insider Threat Community

DPAC-2023-294 □ PERSEREC-RN-23-05 □ NOVEMBER 2023

Michael Hunter, Whitney Fauth, Dean Cisler, & Andrée Rose

Introduction

As data accumulates as a part of U.S. Government Insider Threat (InT) initiatives, fundamental knowledge in research design and statistical principles, along with data governance and analytical capabilities, are required to effectively manage and analyze such data. Data analysis is required to support the Government's Federal Data Strategy (FDS) "to building a robust integrated approach to managing and using data … to generate evidence-based policy, deliver on mission operations, serve the public, and steward resources" (FDS Action Plan, 2021). Likewise, data analysis supports the DoD's data strategy vision of "becoming a data-centric organization that uses data at speed and scale for operational advantage and increased efficiency" (DoD, 2020, p. 1). Additionally, the National Insider Threat Task Force (NITTF) requests data from federal agencies to evaluate annual trends in aggregate to provide insight into the "growth and maturity of the federal government's efforts to identify, deter, and mitigate insider threats" (NITTF, 2021). According to the NITTF "*understanding the challenges they* [Executive Branch departments and agencies] *share in common enables better advocacy for solutions and support to mature and grow the [InT] programs.*"

At the programmatic level, data analysis can be used to evaluate the capabilities, effectiveness, feasibility, and costs of proposed and alternative programs under consideration. To strengthen the InT community's ability to identify and improve prevention and response efforts, the Office of the Under Secretary of Defense for Intelligence and Security funded the Defense Personnel and Security Research Center's Threat Lab to develop a data and analysis Workbook for use by the InT community. The purpose of this Research Note is to document the workbook development process.

Method

To inform Workbook development, the research team launched an outreach campaign to learn about data analysis efforts across InT programs. The team conducted several stakeholder working meetings and semi-structured interviews targeting specific InT programs, which included the Defense Department's Insider Threat Management Analysis Center (DITMAC), Army, Marine Corps (USMC), Navy, Air Force, the Department of State, Defense Counterintelligence and Security Agency, Department of Justice, the National Geospatial-Intelligence Agency, Department of Energy, Federal Reserve, and Office of the Director of National Intelligence. We also deployed a Data Call to

the working meeting participants to ascertain information about reporting practices.

Design and Development of Workbook Content

Feedback from the working group meeting and interviews guided the development of the Workbook content. The team also researched multiple introductory statistics courses and materials, along with relevant research articles and analysis plans published by various government agencies (e.g., the Centers for Disease Control and Prevention). The design of the Workbook was based on instructional strategies and learning objectives to engage end-users to explore data, understand statistical concepts, apply methods, and communicate results. This took the form of partitioning the Workbook into two sections. One section focused on providing the background knowledge and definitions used within statistics, while the other section focused on "hands on" experience in applying the principles and conducting basic analyses using simulated data on "real world" example research questions and data. This structure enabled the application and relevance of statistics to real-world problems within the InT community, rather than just focusing on theory and computation.

Another guiding factor in designing and developing the Workbook was to establish a balance between the breadth and depth of the statistical content covered, considering the broad range of prior knowledge, interests, and end-user needs. This approach required adapting content on statistical principles and analyses to suit different levels of learners, as well as provide additional online and supplemental materials for the end-users. As part of the supplemental material, the team generated an InT-relevant simulated dataset along with the embedded functions to calculate statistics presented in the Workbook's example research questions. This supplemental worksheet includes the functionality for end-users to recreate the results and data visualization capabilities using their constituent data holdings.

Results

Best Practices of Analyses for InT Programs

Developing the analysis Workbook and simulated data was informed by stakeholder meetings, semi-structured interviews, and an instructional strategy that focused on applying statistical principles. A key finding from the outreach campaign and interviews was that most InT programs are at different stages in their analytic capabilities, which also vary in their methodological approach. For example, some InT programs utilize Excel pivot tables and functions to generate dashboards to communicate their findings, other programs utilize predictive and classification techniques using machine-learning approaches.

Interview results revealed several opportunities to enhance data management and data analysis. While some InT programs are starting to conceptualize or develop a centralized data repository, other programs have existing functional repositories. Additionally, some InT programs require direction, or guidance, on reporting operational metrics absence of clear policy for reporting such metrics within their office or Hub, or at the enterprise level. Results from the Data Call on reporting requirements (n = 6; with approximately 20% response rate) showed that 5 out of 6 respondents reported that their corresponding agencies report InT information as required by policy and/or the organization's leadership. Types of information included in their reports were "staffing/resources," "major challenges," "information about insider threats," "mission coverage," and "risk assessment information."

Data Analysis Workbook

Based on information from the outreach campaign and interviews, the Workbook focused on the most basic and essential statistical principles and methodology to address operationally relevant research questions in the InT community. In particular, the learning objectives focused on providing end-users with a basic understanding of

research design and statistical principles. The Workbook offers a framework for thinking about and generating meaningful research questions, developing research questions based on data and conducting basic statistical analyses and data visualization. While this effort was designed for use by any federal government InT program, the Workbook offers familiarity with the DITMAC System of Systems data (DSoS) reporting form (which is the alternate method for reporting; see more details about DSoS in subsequent sections). The Workbook example research questions and analysis focused on categorical analyses, trend analysis, and thematic analysis. The Workbook includes additional information about the functionality of the worksheet in the Appendix.

As described above, the simulated data for this effort was informed by the DSoS reporting form, which is a minimum for reporting and enables aggregating operationally relevant data. For the Workbook, the simulated data were generated based on the variables included in the DSoS form (for instructional purposes), such as submission type, fiscal year, gender, reporting thresholds, and potential risk indicators. A total of 100 simulated data points were generated with statistical patterns preserved (artificial) in the data to visualize, analyze, and interpret for instructional purposes in the Workbook. Integrating the DSoS reporting form further made the Workbook operationally relevant because the example research questions, instructional analyses, and interpretation of results were based on the variables the end-users are currently using in the field. The Workbook should be used in conjunction with the supplemental worksheet. The end-users are encouraged throughout the Workbook to utilize the built-in functions for calculating statistics and data visualization procedures presented in the Workbook.

Conclusions

This effort resulted in an instructional analysis Workbook to support the data management and analysis capabilities within the InT community. The team also created a supplemental datasheet that included operationally relevant variables. This effort also resulted in a better understanding of the opportunities to enhance data management and standardization and identified areas for further data outreach campaigns and instruction to strengthen the InT community's data analytic capabilities, which are outlined below.

Moving Forward with an InT Data Analysis Working Group Toward Data Standardization for the Federal Government

One of the primary findings derived from our meeting with stakeholders across the InT community was a near-universal desire for a standardized approach to data collection. A standardized approach would further ensure that InT programs across the Federal Government are basing decisions on data that are consistently defined, collected, and analyzed in the pursuit of understanding risk factors for outcomes associated with InT reports. The research team strongly advocates a systematic approach to data standardization that begins with establishing an Insider Threat Data Working Group (InT-DWG). The InT-DWG would be made up of representative members from each of the federal InT programs and would be responsible for identifying a core set of variables that are operationally defined and documented in a formal, enterprise-wide data dictionary. Such a data dictionary would serve as the foundational document for data collection and analysis across the Federal Government and would facilitate data sharing, where possible, and cross-domain comparison of case information and outcomes. However, the development of a formalized data dictionary should not preclude the individual agencies and services from collecting and analyzing additional organization-specific metrics. Each individual agency and organization will still be able to collect, analyze, and report on information that meets its unique needs.

The InT-DWG and data dictionary could then form the basis for enterprise-wide data analysis (to include advanced predictive analytics and using expansive data aimed at prevention and intervention) and reporting mechanisms. Organizations would also recognize the benefit of information reciprocity, where necessary and appropriate. As individuals transition between jobs, offices, and organizations, any associated InT data could logically move with them, so providing organizations with longitudinal information and a standardization data management framework

will better inform potential for future risk behavior. The team acknowledges that data governance of data sharing must be clearly defined; some data fields may require redaction, depending on information sensitivity, and that some organizations might be restricted in the information they are allowed to share. The more data we can aggregate, the better able we will be, as an enterprise, to predict, respond to, and prevent serious insider risk events.

Intermediate Workbook and InT Research Design and Statistics Course

This effort also revealed a need to foster the dissemination and instruction of research design and statistical analysis to the InT community. This would result in an intermediate Workbook, which would cover more advanced statistical principles and analysis techniques, including multiple regression, advanced trend analysis, multivariate statistics, imputation of missing data, and machine-learning capabilities. This future effort would be advanced by developing a fast-track research and statistics course designed specifically for the InT community. This course would benefit from utilizing experimental learning and interactive instruction approaches to facilitate the engagement in meaningful real-world analyses and social interaction and communication among participants. This fast-track course would have an accelerated curriculum that focuses on operationally relevant research questions to ensure fast pace to course completion. The course would be led by the research team and tailored for the varying needs across different InT programs and would utilize both the current and intermediate InT analysis Workbooks.

Integration with Prevention, Assistance, Response (PAR) Model for the DoD

DITMAC is at the forefront of establishing PAR capabilities across installations. These centers will enable the local installation to assess and respond to potential risks and InT concerns. The PAR coordinators will receive case-type information that can and should be collected to facilitate understanding of command, local, regional, and branch-wide threat patterns. Currently, there exists no underlying structure to enable PAR efforts to collect data. We recommend that DITMAC promote a common data structure to use across all PAR capabilities. To the extent possible, this standardized data structure should be informed by the InT data dictionary that could be developed by the InT-DWG. This would enable local installations to collect and analyze additional data they are interested in while facilitating data sharing and reciprocity within and across InT hubs and within the broader group of InT organizations. If the data dictionary is implemented early in establishing PAR capabilities, we anticipate a strong methodology to optimally use PAR data that benefit the Federal Government as a whole.

References

Department of Defense. (2020). *DoD data strategy*. https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DOD-DATA-STRATEGY.PDF

Department of Defense, Office of the Inspector General. (2022, September 28). *Audit of the DoD component Insider Threat Reporting to the DoD Insider Threat Management and Analysis Center* (DODIG-2022-141). https://www.dodig.mil/reports.html/Article/3175529/audit-of-the-dod-component-insider-threat-reporting-to-the-dod-insider-threat-m/

Federal Data Strategy Action Plan. (2021). https://strategy.data.gov/assets/docs/2021-Federal-Data-Strategy-Action-Plan.pdf

National Insider Threat Task Force. (2021) *The State of Insider Threat Programs: Trends from Annual Reports, 2018-2020*. [White Paper].