Network Mission Assessment Toolkit

Gabriel Wachman, Steve Ernst
Joshua Haines, Jeff Pattillo

10 August, 2011

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

This work is sponsored by the Department of the Air Force under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Network Mission Assessment Toolkit
Problem Statement, Proposed Solution

CNW, 2009: SAB finds that the "Range of cyber attack effects is not well understood by warfighters" and recommends the development of "technology programs that enhance Mission Assurance"

The Problem:

- **Mission commanders need to understand how a threat to their infrastructure translates to a threat to their mission**
- **Existing risk / vulnerability assessment capabilities can account for mission context, but rely on resource-intensive manual processes**
- **There exists no reliable, automated system for mapping the mission onto the infrastructure**

Our Insight:

- **A mission can be represented by the group of network resources used by mission participants in support of the mission and all the associated network resources on which this group depends**

Our Hypothesis:

- **These mission groups can be inferred automatically from network-derived data using machine learning techniques and limited human contributions**

Our Solution:

- **NMAT: a decision-support tool used to identify mission groups by analyzing network data sources. These groups will be augmented with information such as host criticality, workflow models, and topic information**

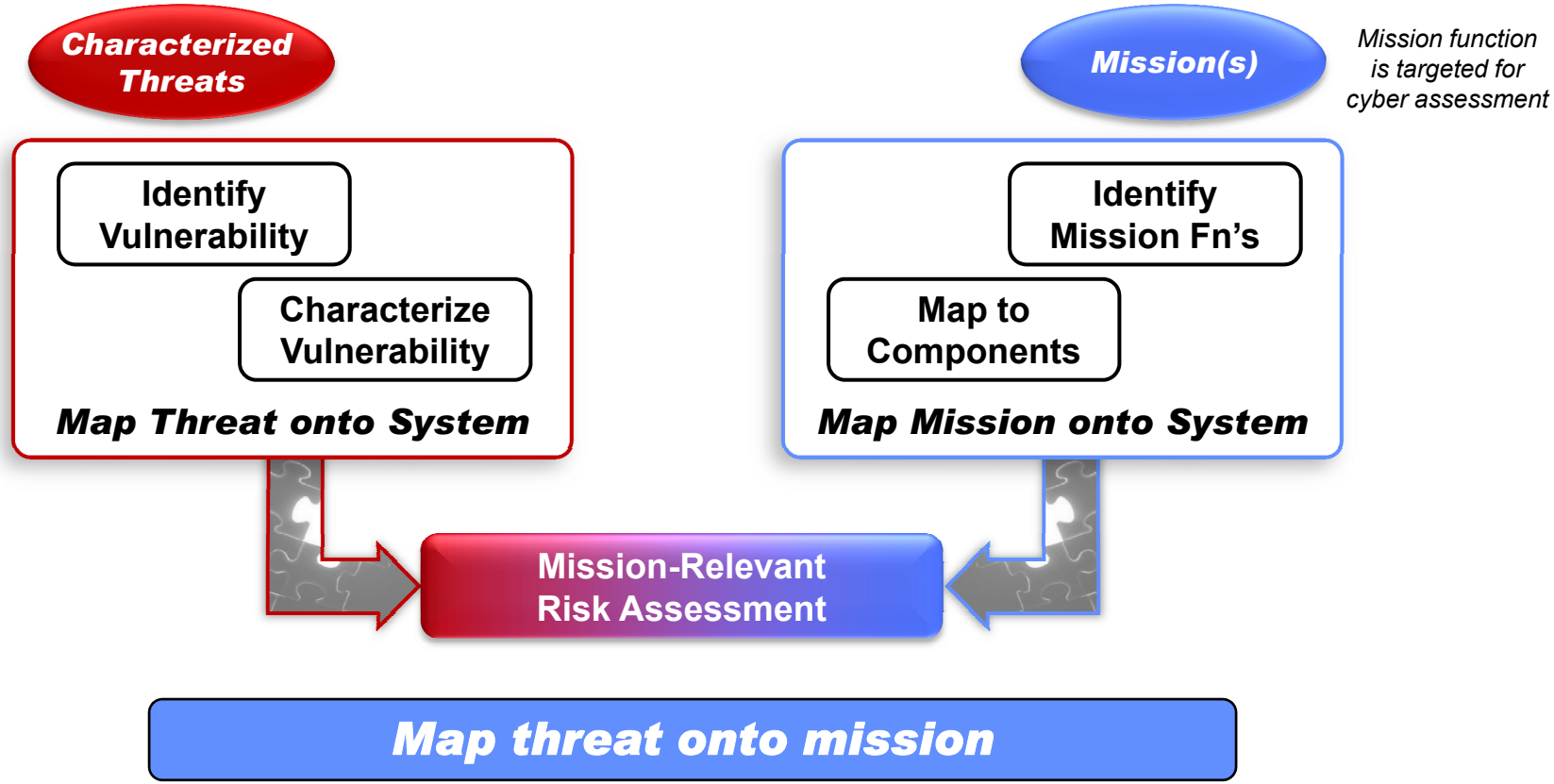
Summary of Mission Assessment Approaches

- NIST: Risk Management Framework (RMF) is a formal specification of the risk assessment process
  - Promotes disciplined, structured and flexible implementation
  - However, is manual process requiring iterative customization of existing baseline models
- MORDA: Formal rigorous red-team methodology for risk assessment
  - Creates complex attack graphs and system models; includes many attack types
  - Allows users to tradeoff security, cost, performance
  - However, time consuming and requires many domain experts
- CNSS-NRAT: Automatic risk analysis for a fixed collection of threats, controls, and network types
  - Provides a high-level risk score and details about threats
  - However, uses high level network and adversary models
- RISKMap: Applies business objectives model to risk via bottom-up approach
  - Identifies infrastructure "crown jewels" that are essential to mission functions
  - Allows exploration of "what-if" scenarios off-line
  - However, requires domain expert inputs
- NMAT: Automated mission-mapping using machine learning techniques
  - Minimal human intervention required; yields repeatable and verifiable results
  - Useful on existing undocumented infrastructure
  - Can be run on an ongoing, periodic basis
  - However, relies upon representation of mission in network traffic

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

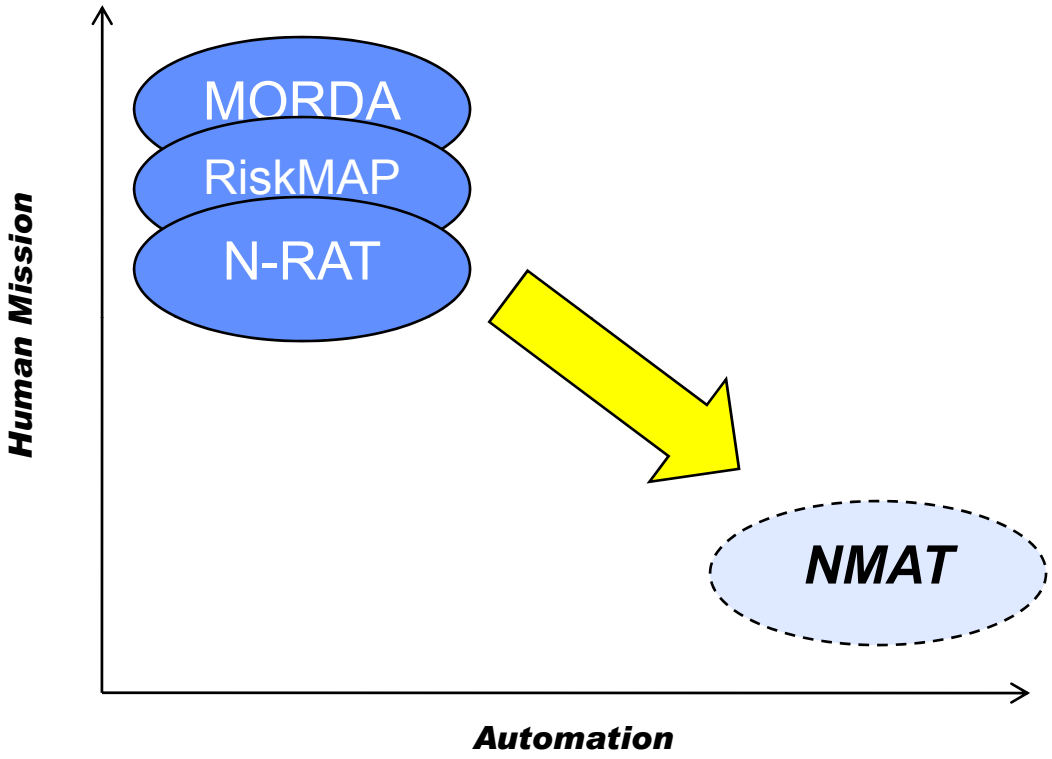
Vision and Approach

Couple with proven threat assessment capability to provide comprehensive risk assessment



1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

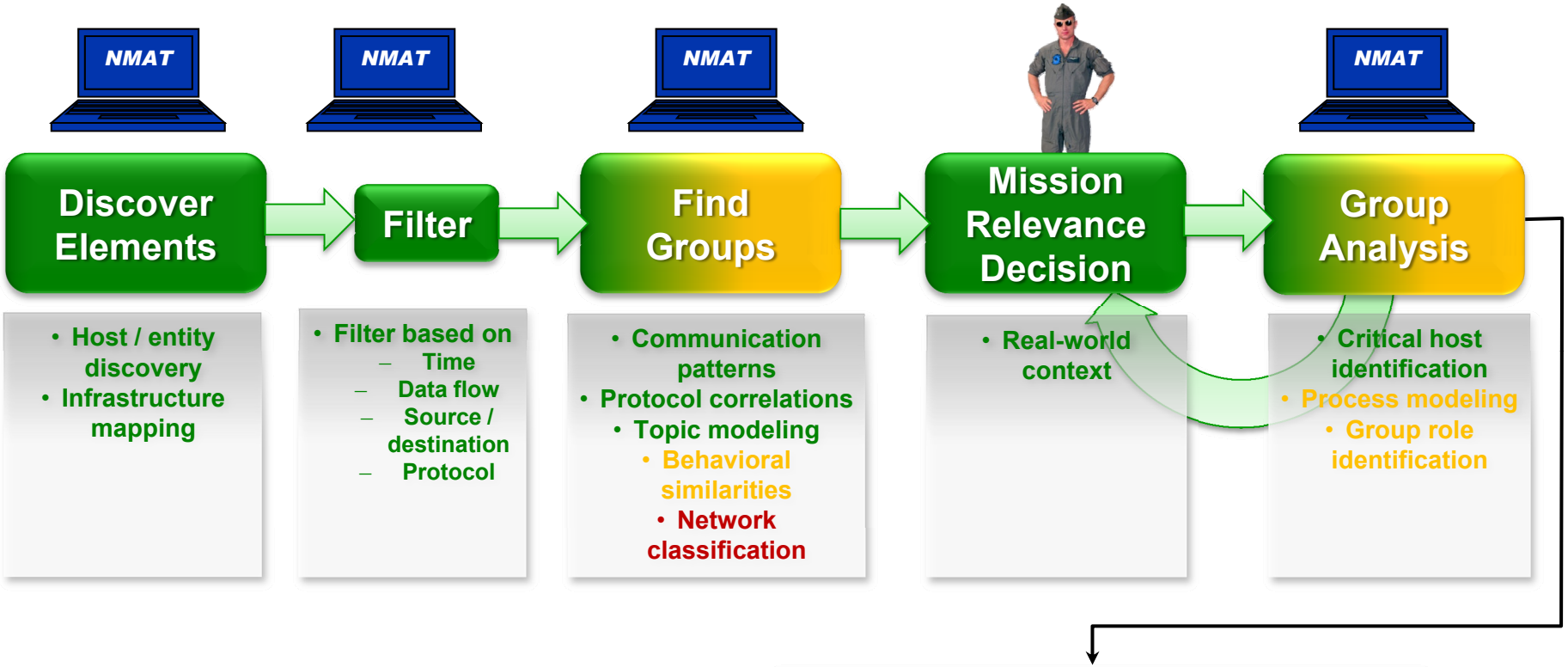
Mission Assessment Approaches



NMAT goal is to be resource-light and automated so that it could be run in "real-time" as mission operations proceed

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

NMAT Processing Pipeline



NMAT Final Output

Mission Model

- Network elements grouped by mission they support
- Associated natural language
- Model of workflow (process)
- Roles of group elements
- Group elements ranked by criticality/importance

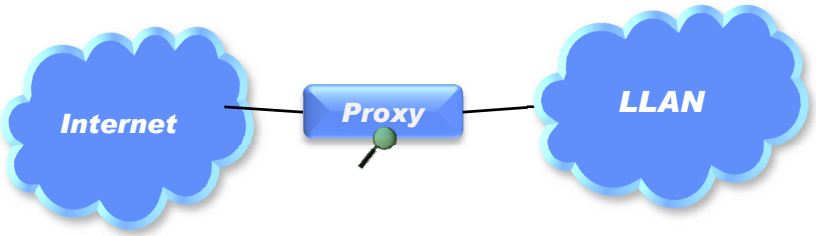
Complete

Continuing Research

Not started

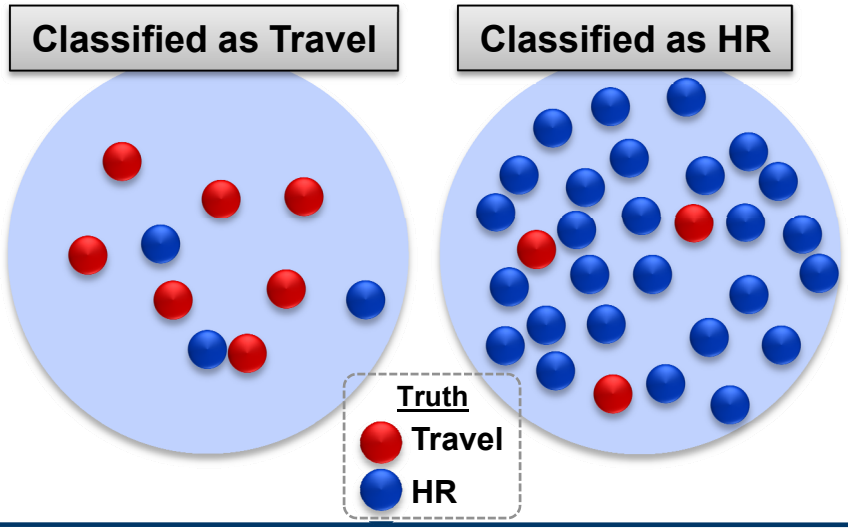
1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LRNOC Data (Identify Groups)



- Proxy logs captured through Splunk
  - All web traffic from the Lab network
  - Limited to 10k records, or about 30 minutes
- Ground truth available through known subnet roles
- Only examine HR and Travel traffic as these mission groups rely heavily on web traffic (HTTP)

- Using J48 decision tree, we can classify known mission participants using behavioral similarity (feature space of length-one labeled walks)
- Classify HR and Travel together with >80% accuracy
- More sophisticated methods should perform much better, but require more data
- Good preliminary results on larger datasets using probabilistic approaches



1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

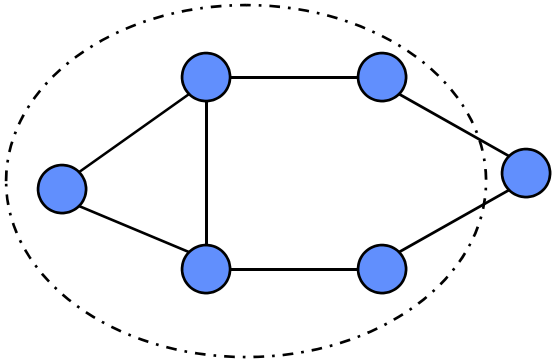
Identify Groups
Probabilistic Approach

- Algorithms from physics, social science
  - Cluster nodes based on node-node communication
  - Nodes that communicate with each other frequently are in the same group
- Mixed Membership Stochastic Blockmodels
  - Airoldi, Blei, Feinberg, Xing 2008
  - Nodes represent IP addresses
    - Each node draws a group membership vector from a Dirichlet distribution
    - Every pair of nodes draws membership from a multinomial parameterized by their membership vectors
    - Edges are drawn from a Bernoulli distribution parameterized by the group memberships of the nodes
  - Exact solution is intractable but good approximation schemes exist (e.g., variational inference, Gibbs sampling, MCMC).

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Identify Groups
Graph Theoretic Approach

- **Clique relaxations**
  - Flexible sub-structure that can indicate community
  - Excellent performance on public social-network data
- **K-clique**
  - Subset of vertices such that the shortest path between every pair of vertices is of length at most K
- **P-connected K-clique**
  - Subset of vertices such that there are K edge-distinct length P paths between every pair of vertices

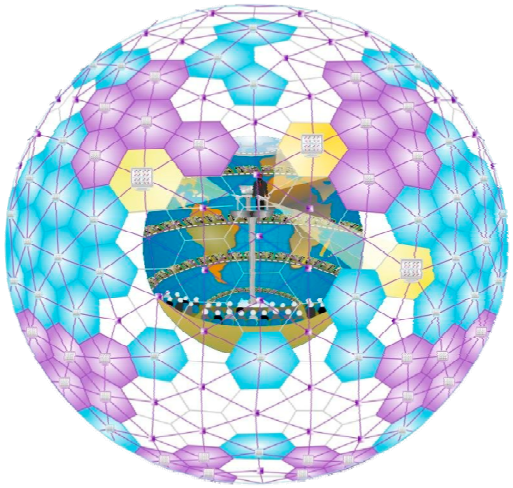


2-clique

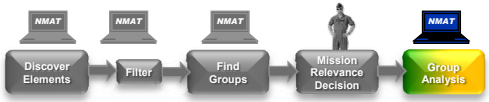
1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Internal Mission Data (Topic Discovery, Process Mining)

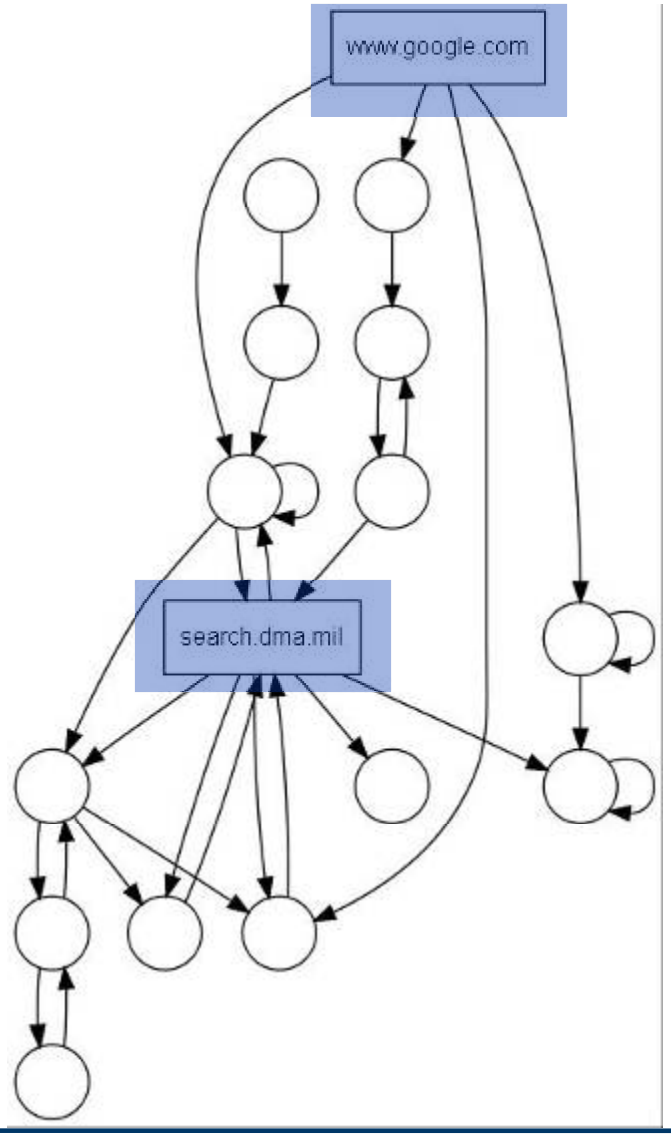
- Mission: Five people gather information about air force bases
  - Find 5 air force bases each
  - Find their commanders, location, number of resident personnel
  - Send information via e-mail to one person who combines all results
  - Information posted to internal web page over interactive SSH session
  - All participants notified of web page update
- All mission activity interspersed with non-mission activity
  - Check e-mail, chat with friends
- Each participant captured network traffic from their machine



Internal Mission Analysis: Process Mining Output



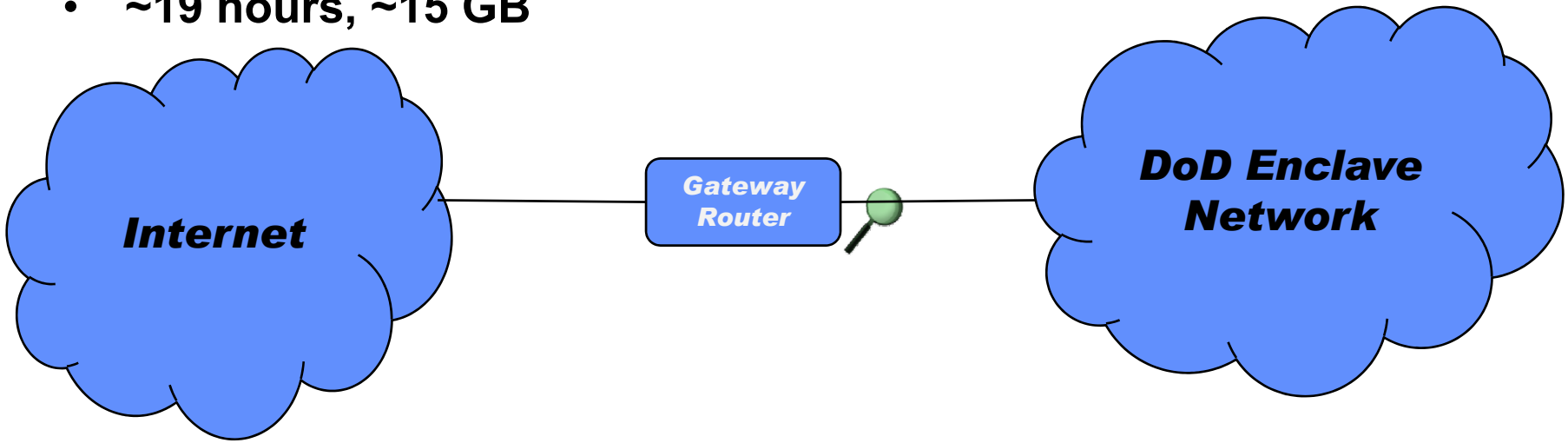
- First concentrate only on HTTP traffic from one user
- Manually identify Google as a starting point
  - This makes sense in search, explore, search template
- Eliminate obvious noise
- Process analysis using PROM reveals model of how a repeated task is executed
- Further graph mining (e.g., PageRank, min cut) can rank nodes, identify failure points



1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

DoD Enclave Data (Critical Host Identification)

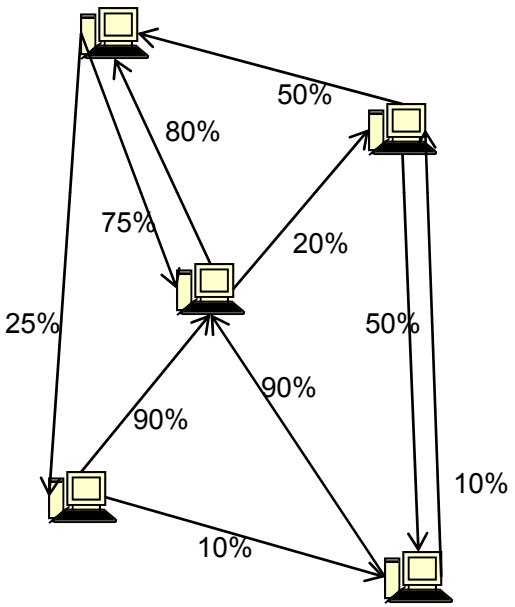
- High volume, multiple mission network
- Network tap at border gateway
  - Regular business hours, no known special events
- Limited amount of network information available
  - Requires constant communication with administrators
- ~19 hours, ~15 GB



1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

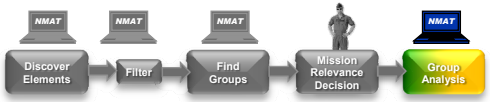
Critical Host Identification

- Model network as a Markov Chain
- States are IP addresses
- Transition probability
  - log(number_of_bytes)
- Critical hosts have higher probability in stationary distribution
- Identifies
  - Critical hosts
  - Non-mission traffic (noise)



1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

DoD Enclave Data Analysis: Critical Host Identification



Critical hosts identified

- X.X.X.A (leave reservation, end of November)
- X.X.X.B (information portal)
- X.X.X.C (information portal)
- disa.mil
- microsoft.com
- msftncsi.com

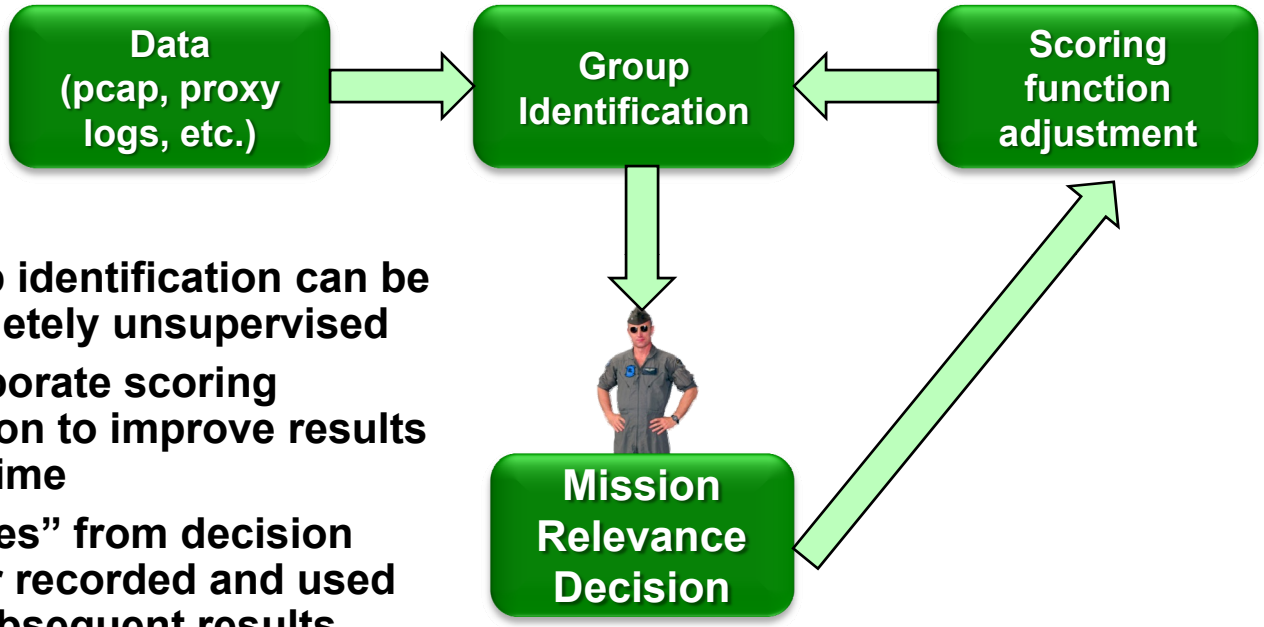
1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Deployable Features

- Several modules could be deployed soon in a decision support role
  - Critical Host Identification
    - Help operator prioritize issues and attention
  - Group Identification
    - Can be used on almost any data source (PCAP, proxy logs, server logs, e-mail logs, etc.)
    - Flexible "target group" design
    - "False positive" rate can be reduced over time as ground truth is given
  - Topic Modeling
    - Works almost off the shelf
    - Requires natural language
    - Best used as a way to enrich other modules

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Training the Modules



- Group identification can be completely unsupervised
- Incorporate scoring function to improve results over time
- "Scores" from decision maker recorded and used for subsequent results

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Conclusion

- NMAT applies various techniques in a new setting
  - Probabilistic approaches to graph analysis
  - Structural approaches to graph analysis
  - Stationary distribution of a Markov Chain
  - Process modeling
- Several steps in the pipeline deployable individually
  - Critical host id
  - Group id
- Offers key analysis in a decision support role
  - Shows operator how mission is supported by the infrastructure
  - Enables operator to focus on most important issues