Activity-Based Community Detection

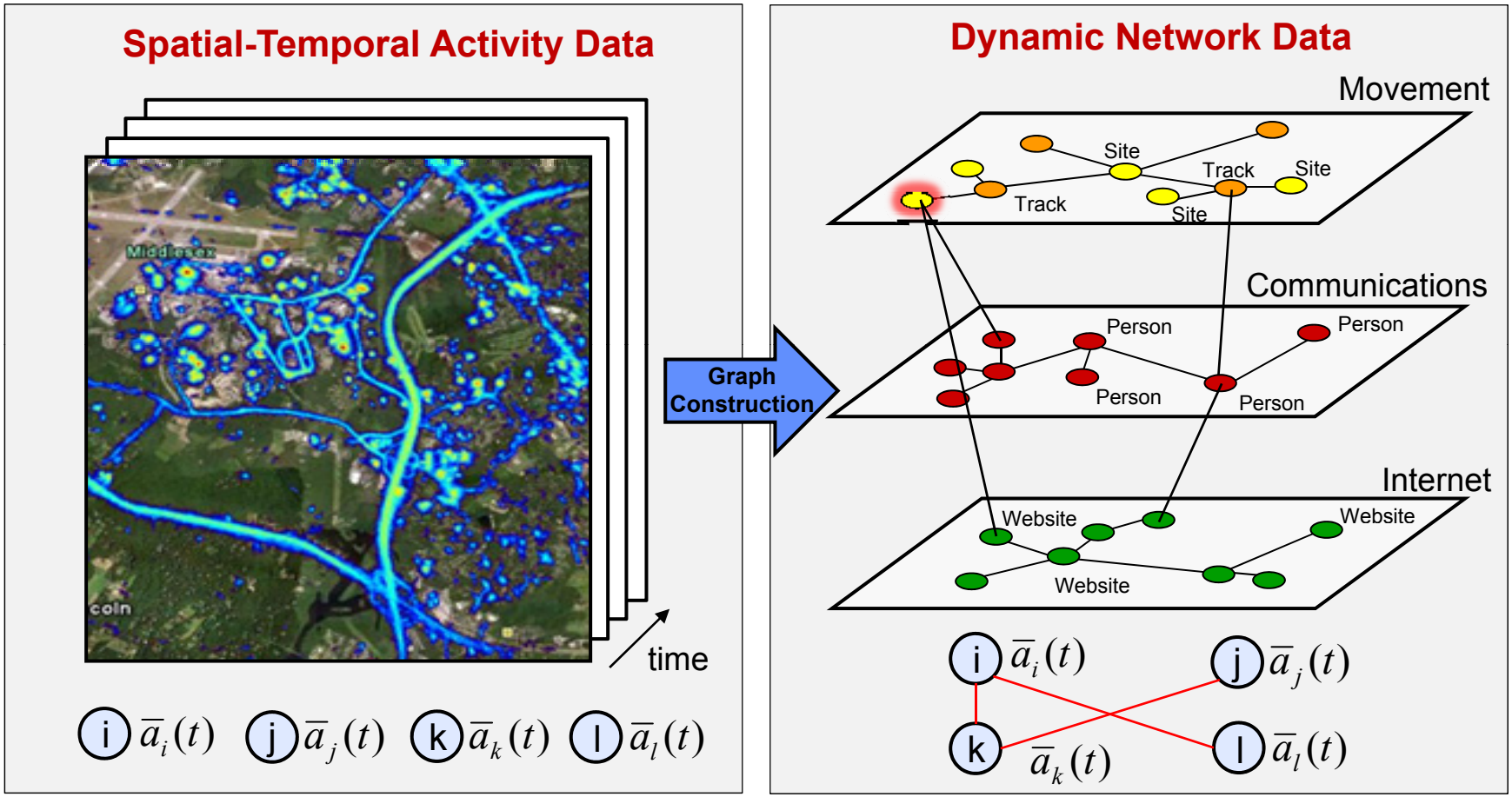
Scott Philips, Edward Kao,
Michael Yee and Christian Anderson

Graph Exploitation Symposium
August 9th 2011

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

This work is sponsored by the Office of Naval Research 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.

Adding Structure to Space-Time Data



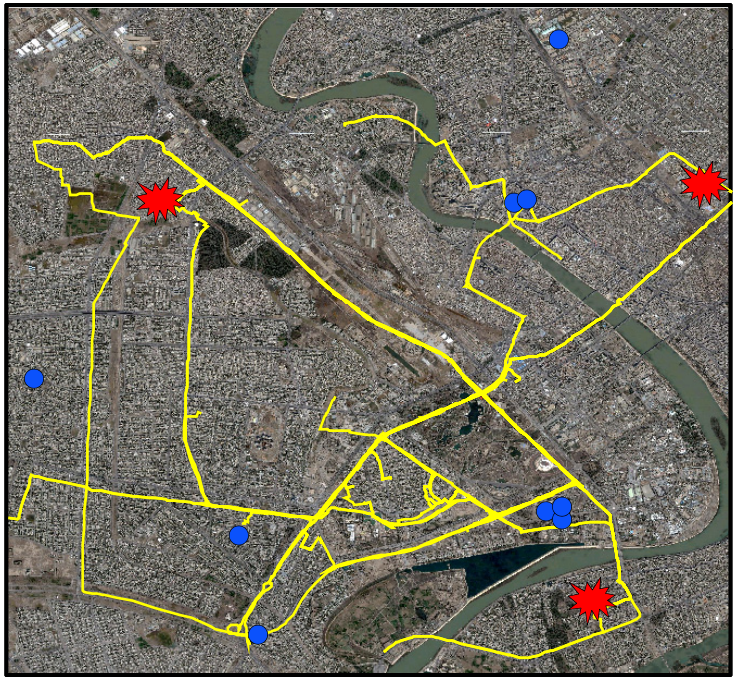
Adding structural information to spatial-temporal data enables higher-level, cross-domain understanding

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Motivating Example: Wide Area Motion Imagery (WAMI)

How can we detect threatening people/places/things and then use that information to prioritize human effort

Imagery + Tracks + Multi-INT Detections



= Multi-INT Detections

Focus Points

Algorithmic Assistance

Where in space and time should I devote resources?

$$P_{\text{Site}}\left(\text{Threat} \mid \text{Data}\right)$$

Which vehicle should I be following?

$$P_{\text{Track}}\left(\text{Threat} \mid \text{Site}\right)$$

Does the vehicle track make a good association?

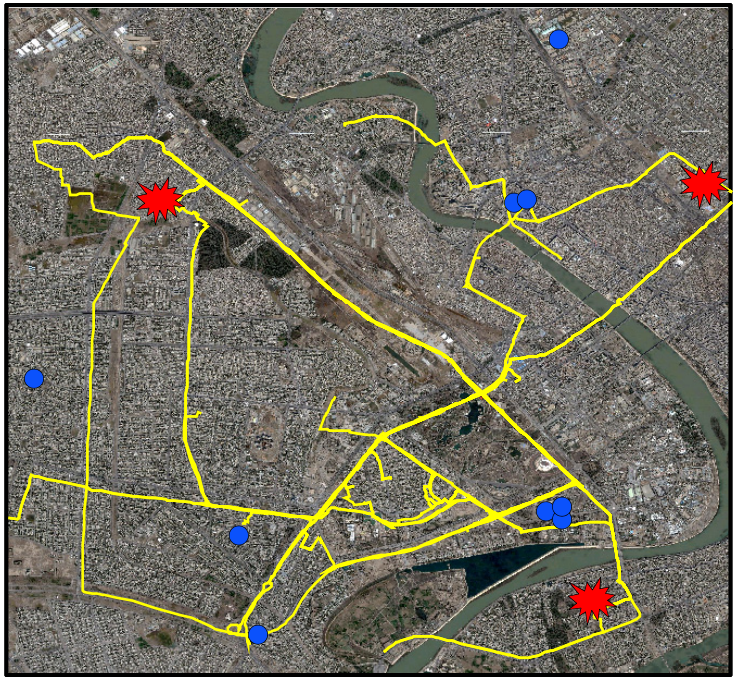
$$P\left(\text{Break} \mid \text{Track}\right)$$

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Motivating Example: Wide Area Motion Imagery (WAMI)

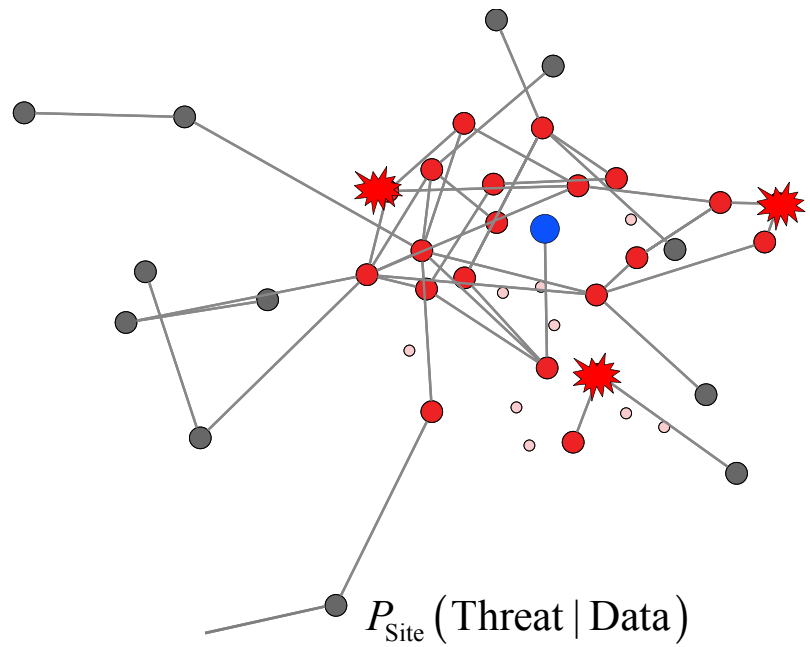
How can we detect threatening people/places/things and then use that information to prioritize human effort

Imagery + Tracks + Multi-INT Detections



= Multi-INT Detections

Graph Abstraction



$P_{\text{Site}}(\text{Threat} \mid \text{Data})$

$P_{\text{Track}}(\text{Threat} \mid \text{Site})$

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

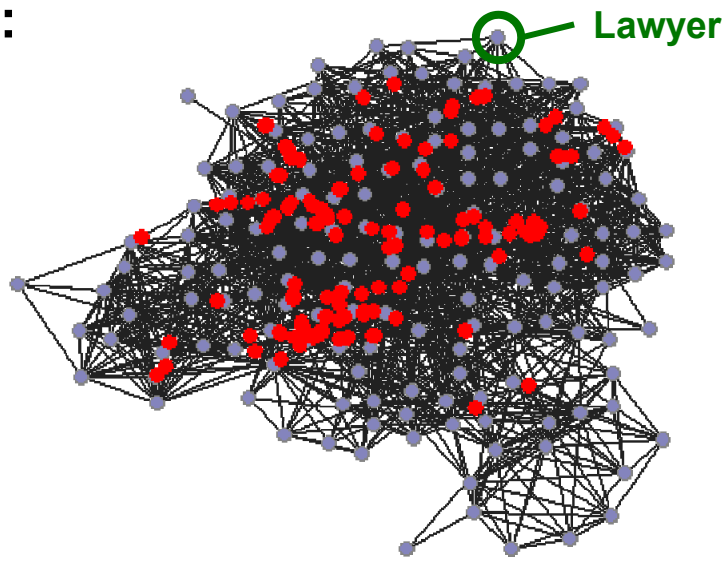
Cued Community Detection on Graphs

Goal: Detection of hidden communities embedded in a large background populations

Approach: Cued community detection algorithms on dynamic graphs

Graph representation of a social network:

- **Network Topology:**
  - Set of all possible associations
  - Provides the substrate upon which interactions occur
- **Network Dynamics:**
  - Local interactions throughout an observation period



Enron Dataset

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Outline

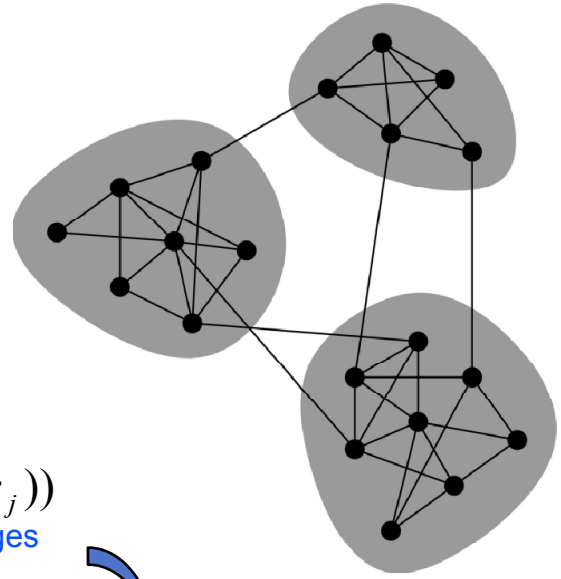
- **Network Community Detection**
  - Background
  - Activity-Based Dynamics
- **Data and Results**
  - Simulated Data
  - MOVINT Data
  - Email Data
- **Network Exploration**
  - Methodology
  - Initial Results
- **Summary and Future Work**

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Community Detection Background:
Global Methods

- **Homophily assumption: Detection of densely connected groups of nodes, with only sparse connections between groups**

- **Edge counting techniques:**



Hamiltonian minimization (Reichardt and Bornholdt)

$$\arg\min_{C} H(C) = \sum_{\{i,j\}\in V} \underbrace{-a\, A_{ij}\delta(c_i,c_j)}_{\text{within group edges}} + \underbrace{b\,(1-A_{ij})\delta(c_i,c_j)}_{\text{within group non-edges}}$$
$$+ \underbrace{c\, A_{ij}(1-\delta(c_i,c_j))}_{\text{between group edges}} - \underbrace{d\,(1-A_{ij})(1-\delta(c_i,c_j))}_{\text{between group non-edges}}$$

Simplifying assumptions

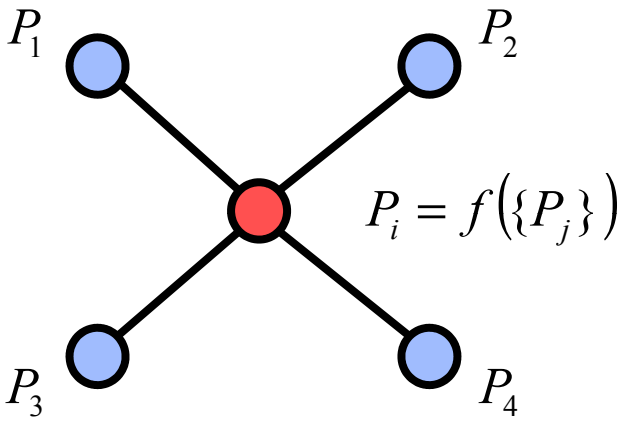
Modularity maximization (Newman)

$$\arg\max_{C} Q = -H(C) = \sum_{\{i,j\}\in V} (A_{ij} - \frac{k_i k_j}{2m})\delta(c_i,c_j)$$

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Community Detection Background:
Local Methods

- **These algorithms iteratively update each node's attributes based on the attributes of its neighbors**



- **Examples include:**
  - **Local modularity (Clauset)**
  - **Label propagation (Raghavan et al. , Gregory)**
  - **HITS* (Kleinberg)**

Still based on homophily: members of the same community are more likely to interact

*Hyperlink-Induced Topic Search

Membership Propagation

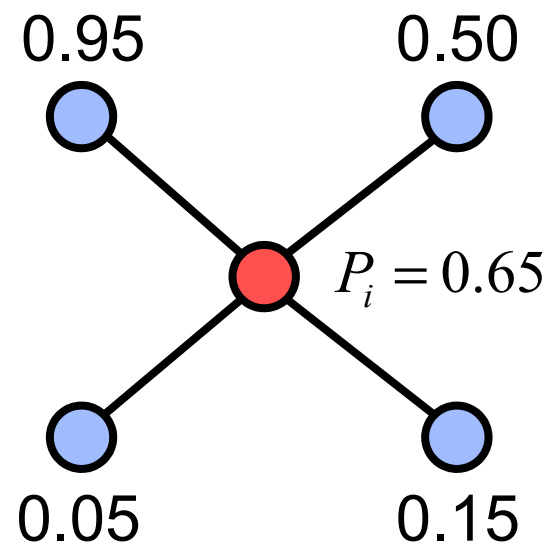
Inferring community membership using local learning

Estimate a node's membership based neighboring nodes

- **Inspired by HITS* (Kleinberg)**
- **Diffusion-like iterative updates**

$$P_i = f\left(\{P_j\}\right) \quad \text{where} \quad j \in N(i)$$

$$P_i = \alpha\left(\underbrace{\frac{\lambda}{|N(i)|}\sum_{j\in N(i)} P_j}_{\text{Average neighbor membership}} + \underbrace{(1-\lambda)\max_{j\in N(i)} P_j}_{\text{Maximum neighbor membership}}\right)$$



*Hyperlink-Induced Topic Search

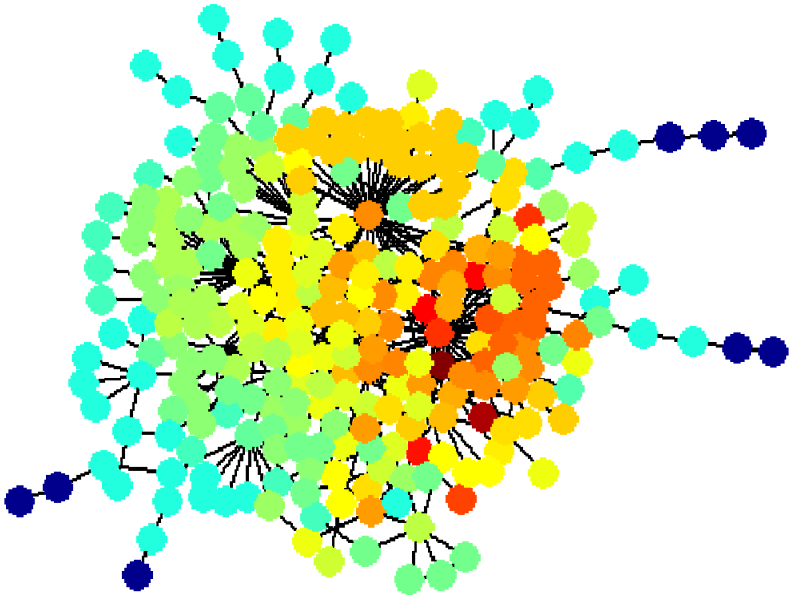
1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Membership Propagation

Inferring community membership using local learning

Estimate a node's membership based neighboring nodes

- **Inspired by HITS* (Kleinberg)**
- **Diffusion-like iterative updates**



*Hyperlink-Induced Topic Search

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Activity-Based Community Detection

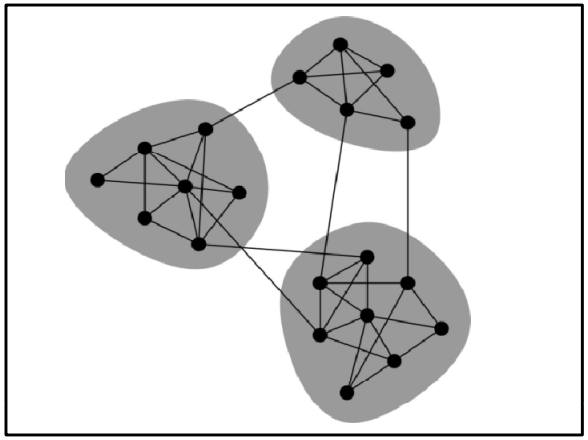
Real world social networks:

- Communities are diffuse (*i.e.* not tightly connected)
- People act on behalf of different communities at different times

Affiliation-Based Community Detection

Densely connected groups of nodes, with sparser connections between groups

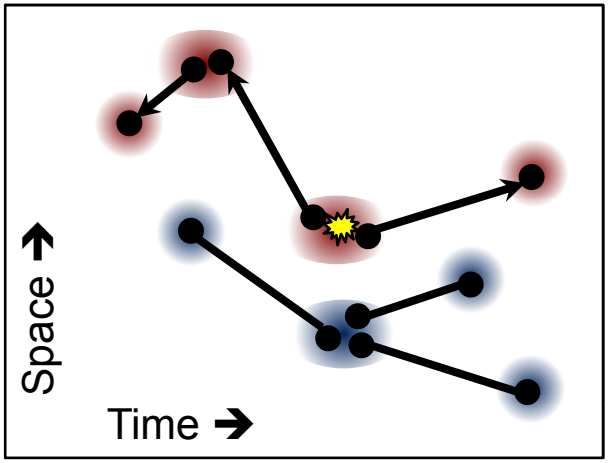
Static Graph



Activity-Based Community Detection

Subset of nodes engaging in an coordinated sequence of finite-duration activities

Space-Time Graph

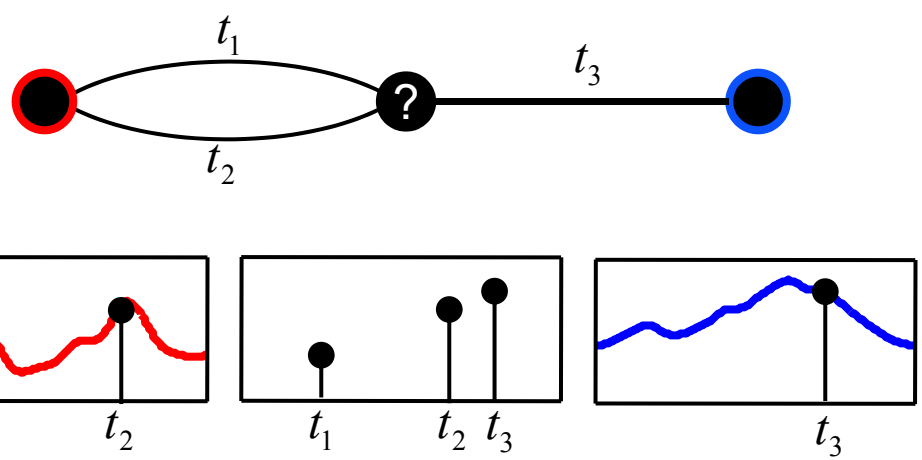


1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Dynamic Membership Potential

- **A nodes potential to act as a members of a specific community can vary over time**
  - *e.g.* Between 9-5 I have my Lincoln hat on
  - This is similar to the role vector in Mixed Membership Blockmodel

- **Membership probability now propagates depending a nodes role at the time of the interaction**
  - The challenge is to estimate role in between interactions



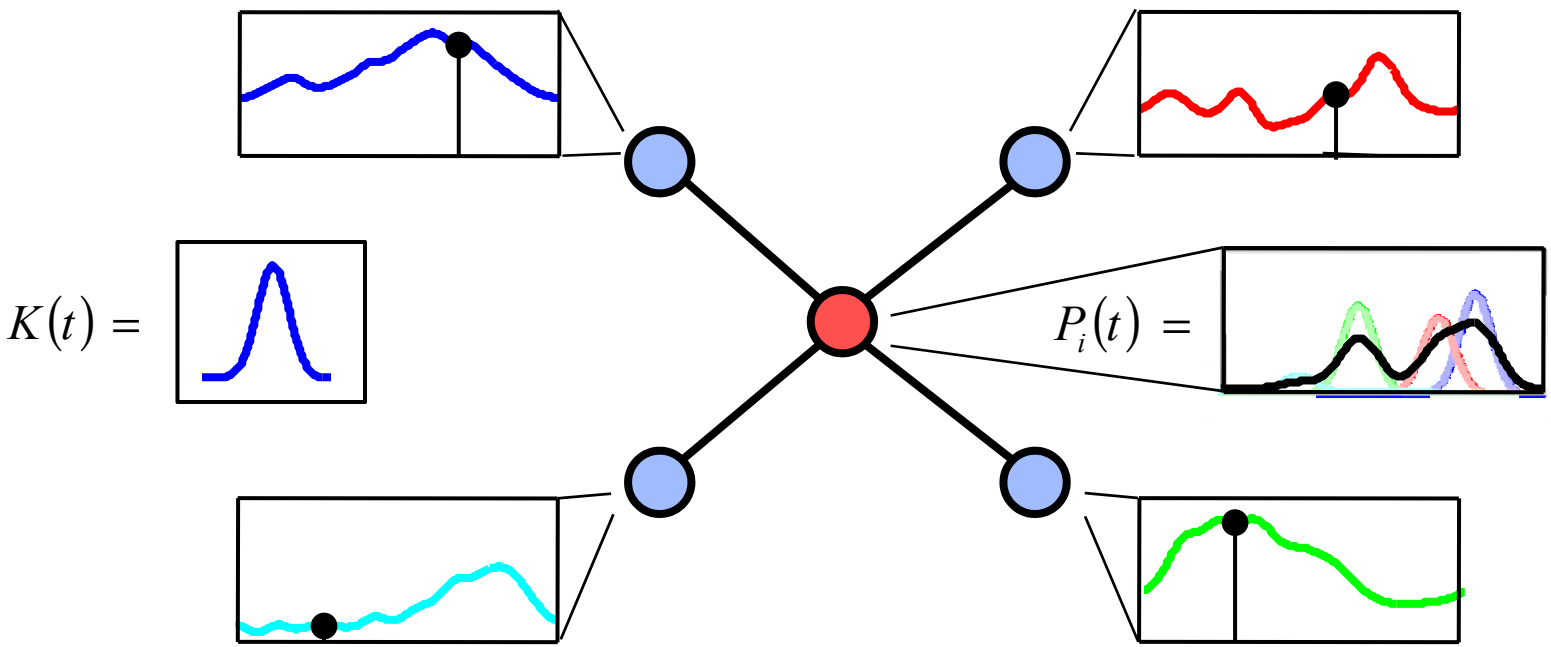
1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Dynamic Membership Propagation

Detection of nodes acting in a coordinated manner over time

Estimate time-varying membership based on edge times

- **Define a kernel function matched to expected community behavior (e.g. co-occurrence of interactions)**



1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Dynamic Membership Propagation

Detection of nodes acting in a coordinated manner over time

Estimate time-varying membership based on edge times

- **Define a kernel function matched to expected community behavior (e.g. co-occurrence of interactions)**

$$P_{s_i}(t) = \alpha_s \left(\underbrace{\frac{\lambda}{|E(s_i)|} \sum_{e_{iv} \in E(s_i)} P_{S_v}(t_{e_{iv}}) K(t - t_{e_{iv}})}_{\text{Average neighbor probability}} + \underbrace{(1-\lambda) \max_{e_{iv} \in E(s_i)} P_{S_v}(t_{e_{iv}}) K(t - t_{e_{iv}})}_{\text{Maximum neighbor probability}} \right)$$

where…

$E(s_i)$ = Set of edges connected to s_i

e_{iv} = An edge between s_i and s_v

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Outline

- Network Community Detection
  - Background
  - Activity-Based Dynamics
- **Data and Results**
  - Simulated Data
  - MOVINT Data
  - Email Data
- **Network Exploration**
  - Methodology
  - Initial Results
- **Summary and Future Work**

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Simulated Graph Data

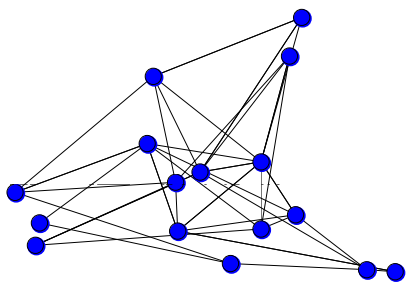
Concept: Generate the topology independently of the dynamics

- **Generate topology using static generators**
- **Generate a dynamic process on top of the topology**

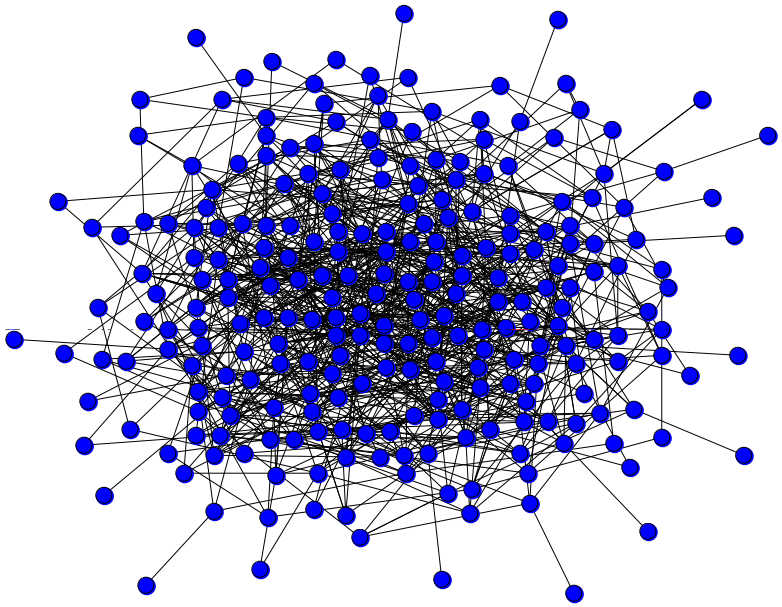
  Foreground – Coordinated Message Passing

  Background – Random Connection Timing

Foreground Graph



Background Graph

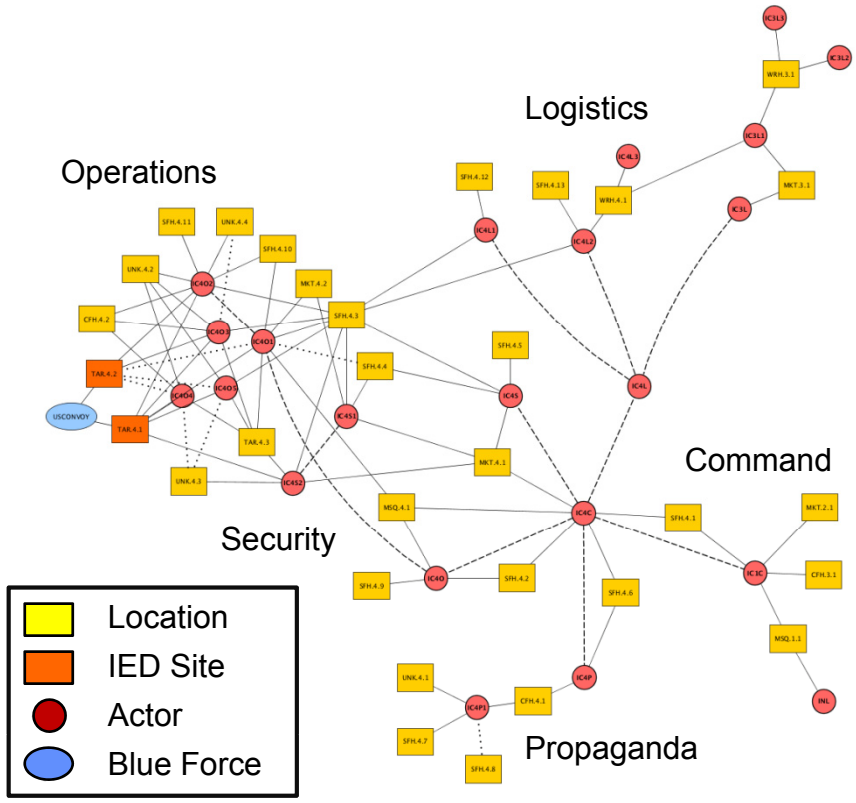


1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

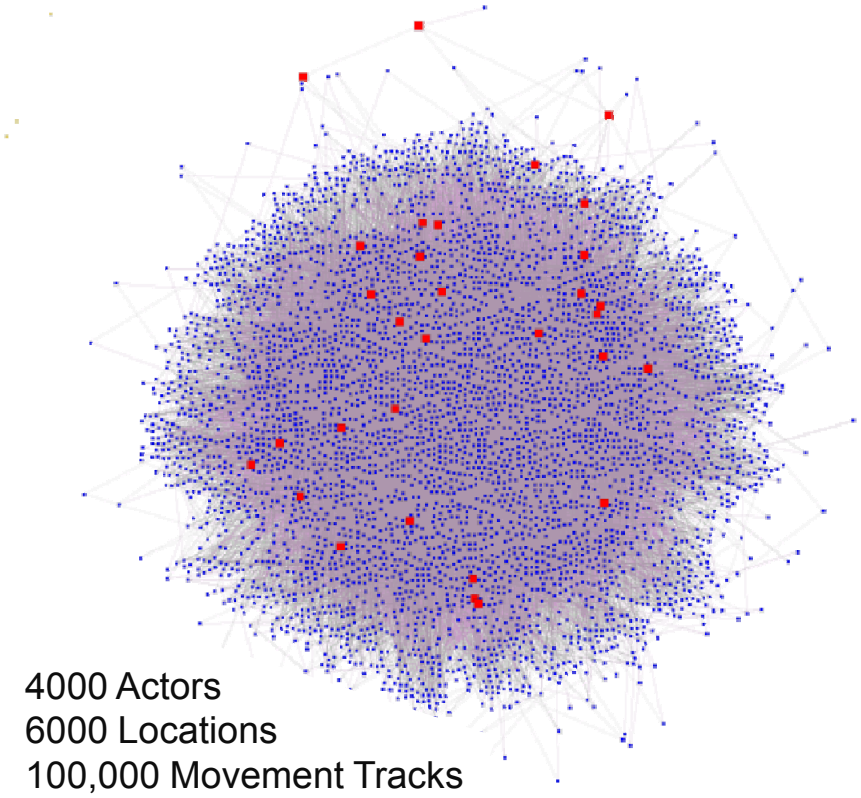
MOVINT Graph Data

Simulated insurgent network hierarchy embedding in a large civilian traffic in Baghdad

NGA Threat Scenario: IED Attack



IDA* Background Model: Civilian Traffic in Baghdad



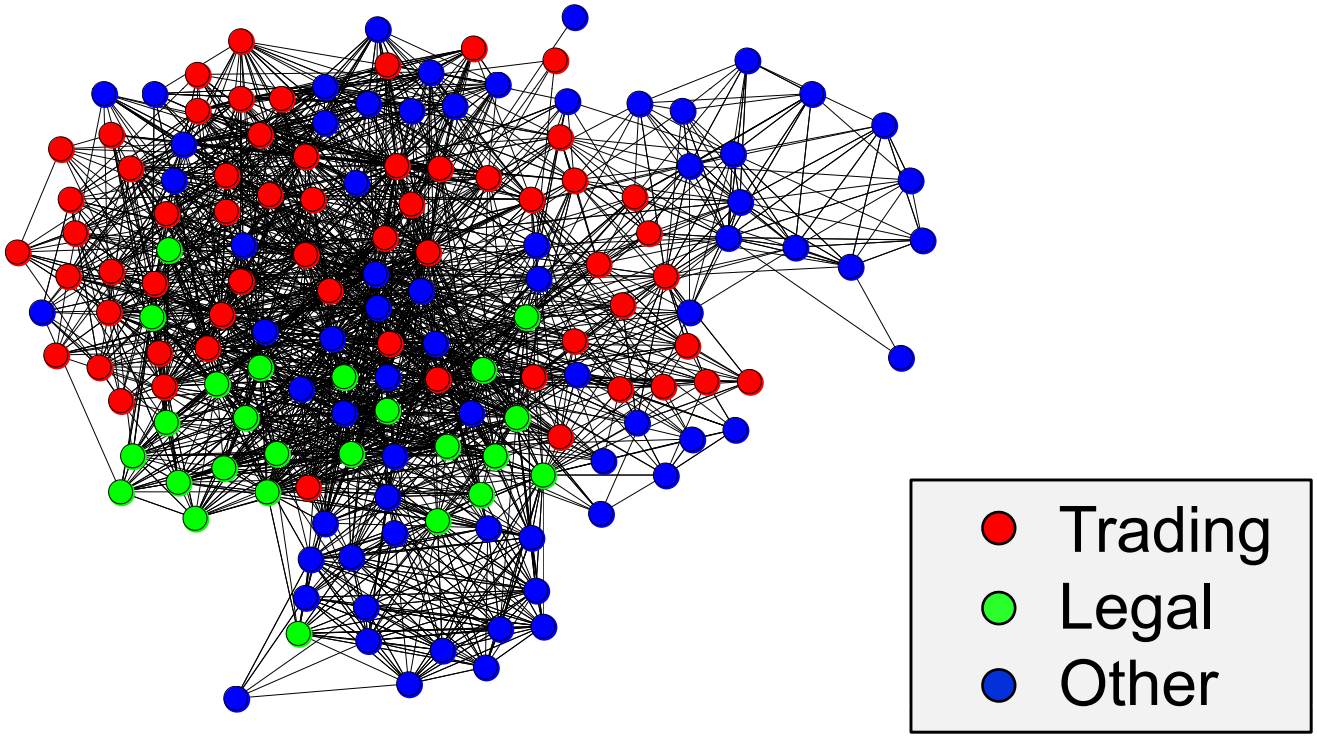
4000 Actors
6000 Locations
100,000 Movement Tracks

*Institute for Defense Analysis

Email Graph Data

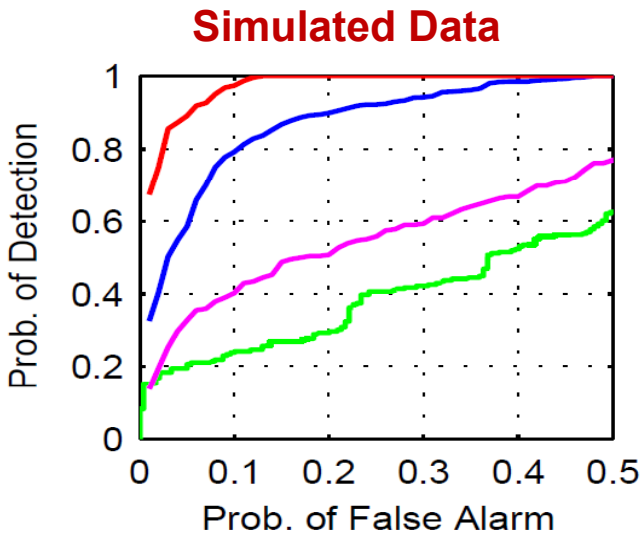
Enron email dataset from Patrick Perry

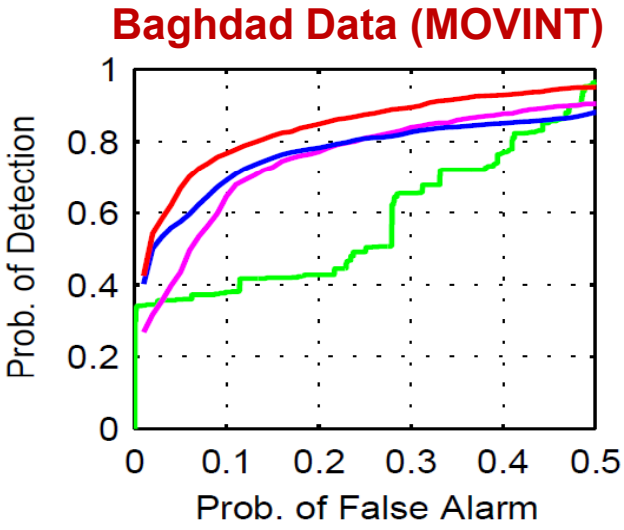
- Nodes represent email accounts for higher management
- Edges represent emails sent between people
- Coloring is by department

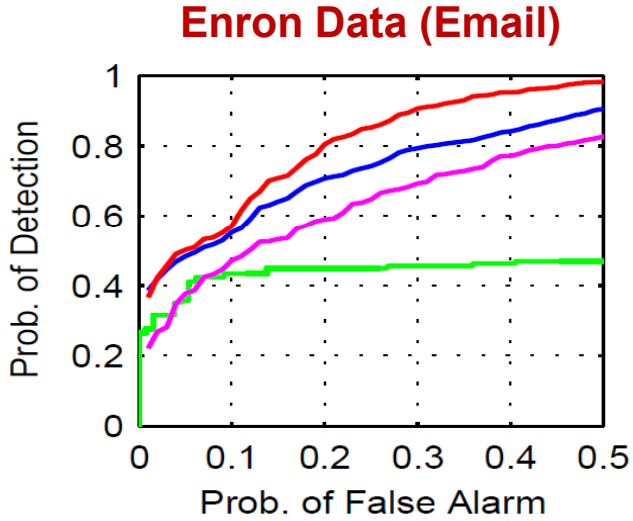


1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Community Detection Results







- Spectral
- Local Modularity
- Static TP
- Dynamic TP

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Outline

- **Network Community Detection**
  - Background
  - Activity-Based Dynamics
- **Data and Results**
  - Simulated Data
  - MOVINT Data
  - Email Data
- **Network Exploration**
  - Methodology
  - Initial Results
- **Summary and Future Work**

Network Discovery
Workload Prioritization

- **The amount of data for an analyst to look at grows exponentially**
  - **Following only a few vehicles from one sensor can lead to hundreds of possible new sites and vehicles to investigate**
  - **This problem affects all aspects of data forensics**

- **Network Exploration Workflow**
  - **Given a set of information, what evidence do I collect next?**

Methodology

1. Propagate membership across network
2. Identify nodes with maximal membership
3. Identify edges with maximal membership
4. Grow graph
5. Repeat

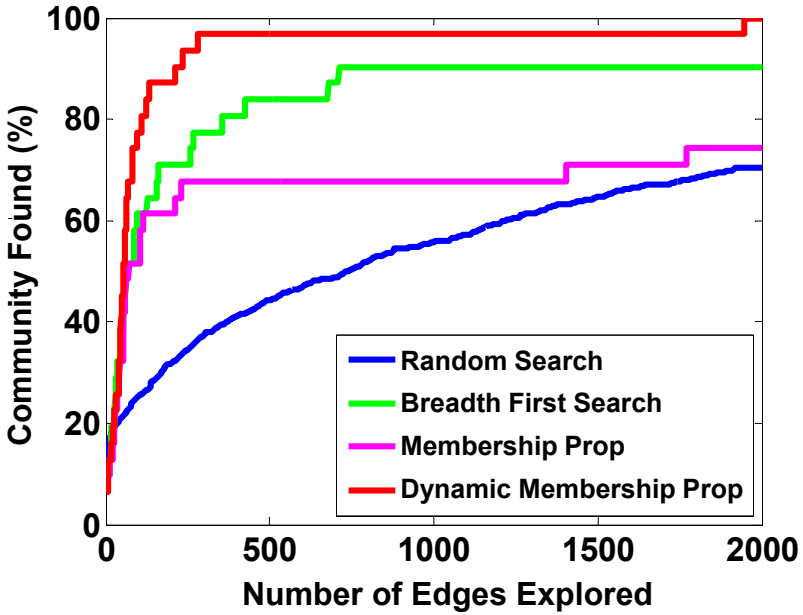
1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Network Discovery
Workload Prioritization

Prioritization Strategies:

- **Random Search**
  Explore tracks randomly starting from a tip node
- **Breath First Search**
  Explores nodes closest to the tip node with lowest degree first
- **Membership Prop**
  Explores nodes with highest threat first
- **Dynamic Membership Prop**
  Explores tracks with highest threat first

Edges Examined vs. Community Found



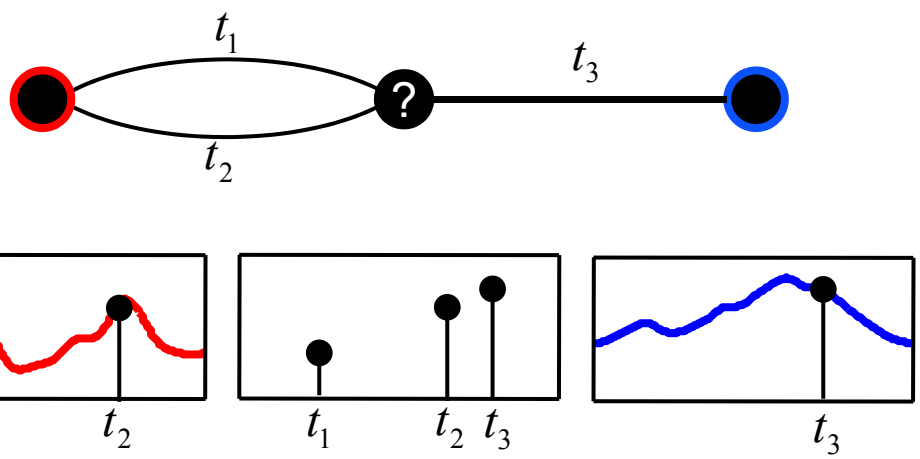
Prioritization discovers 87% of the insurgent network using only 0.5% of the total potential edges.

1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Summary

- **Static structure alone is often insufficient to accurately characterize social groups.**

- **Activity-based communities are defined by a set of nodes engaging in an a coordinated sequence activities**
  - Disease and rumor spreading
  - Collaboration between colleagues
  - Coordinated activities between insurgent cells



1951–2011
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY