

Towards a Detection Theory For Intelligence Applications

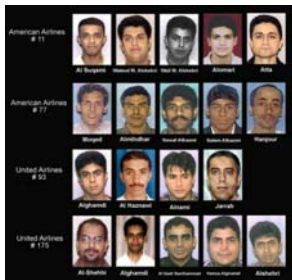
Stephen Ahearn
Jim Ferry
Darren Lo
Aaron Phillips

4th Conference on **Mathematical Methods in Counterterrorism**

September 20-22, 2007
Rochester, NY

Motivation

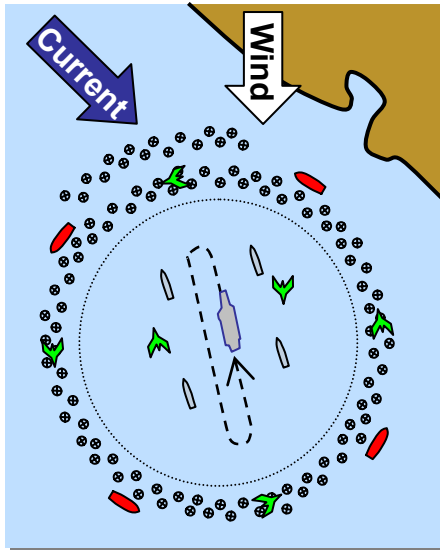
- Classical detection theory has been developed and applied for decades to detect and track stealthy targets
 - Submarines, aircraft, missiles, ...
 - Using a variety of sensors: sonar, radar, IR, ...
 - Gives U.S. a distinct advantage over adversaries



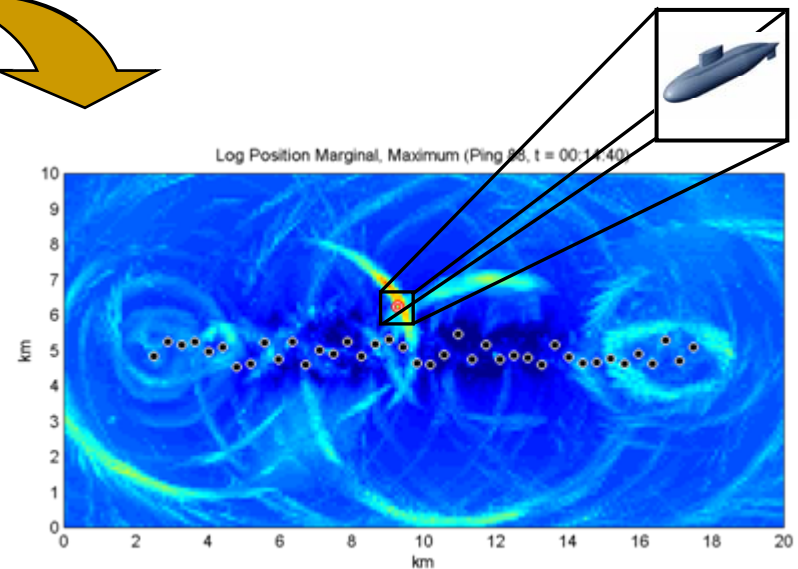
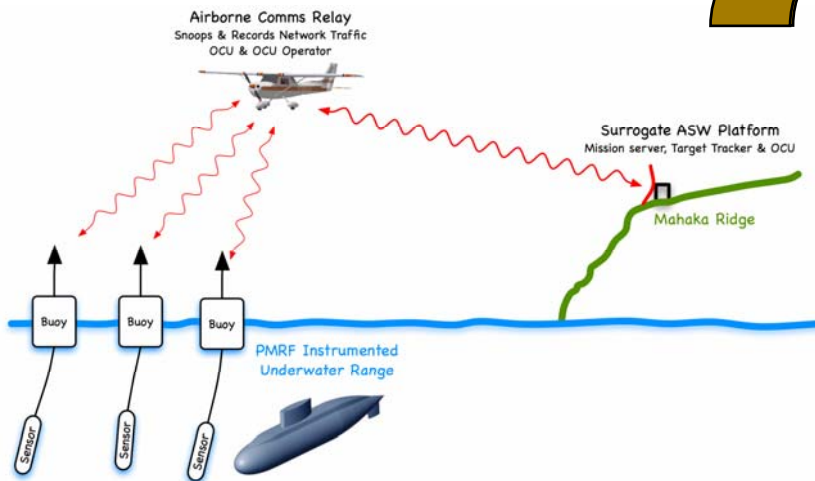
- We seek to develop an analogous theory of detection on networks
 - To detect and track threat networks and activities that cannot be observed directly
 - Exploiting diverse data: SIGINT, HUMINT, IMINT ...
 - To maintain our advantage over today's adversary and win the GWOT



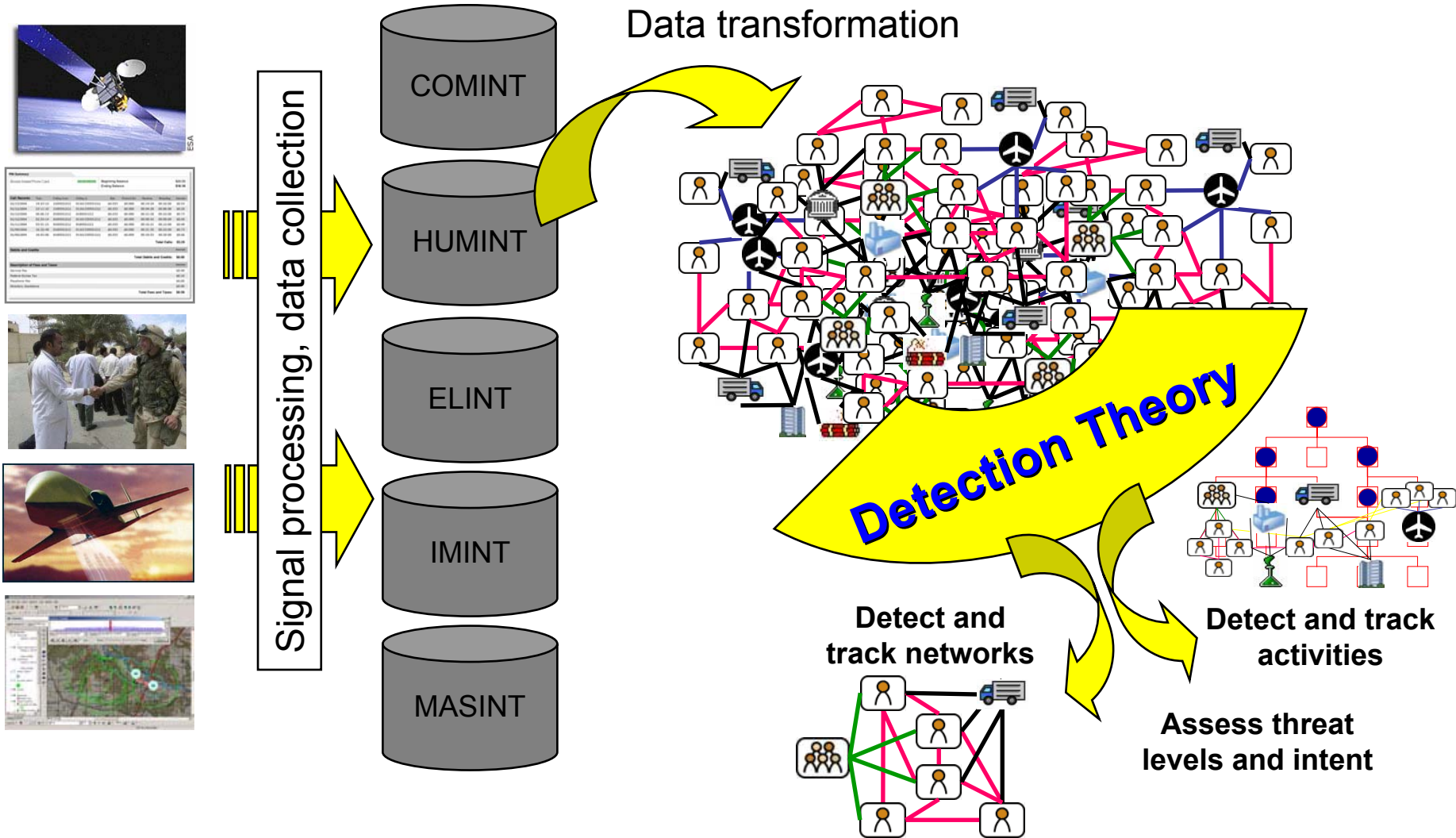
Classical Detection Theory Applications



- Detection theory enables detection and tracking of stealthy targets
 - Such targets cannot be detected by analyzing sensor reports separately
 - Only through principled data fusion does the signal stand out from noise
 - The probabilistic framework correctly manages uncertainty and risk



Detection Theory for Intelligence Applications



Approach: Two well-established theories

- Likelihood Ratio Detection and Tracking
 - Explicitly models noise and signal
 - Principled Bayesian framework for managing uncertainty
 - Used for decades for tracking stealthy kinematical targets
 - Traditional domain has metric structure
- Random graph theory
 - Models transactional domain
 - Discrete structure
 - Rich mathematics

Likelihood Ratio Detection and Tracking (LRDT)

■ Requirements

- State space $\mathbf{X} = \tilde{\mathbf{X}} \cup \{\emptyset\}$
- Measurement model $L(z | x)$ (or $\mathcal{L}(z | x) = L(z | x) / L(z | \emptyset)$)
- Motion model $P^T(x_t | x_{t-\Delta t})$

■ Result

- Update equations: probability form — $P(x_t)$

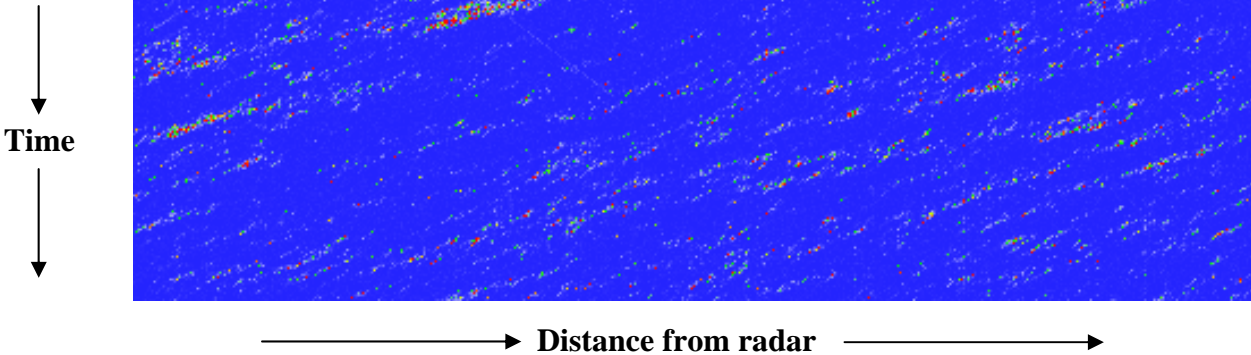
$$P^-(x_t) = \int P^T(x_t | x_{t-\Delta t}) P(x_{t-\Delta t}) dx_{t-\Delta t} \quad P(x_t) = \frac{1}{C} L(z_t | x_t) P^-(x_t)$$

- Update equations: likelihood ratio form — $\Lambda(x_t) = P(x_t) / P(\emptyset)$

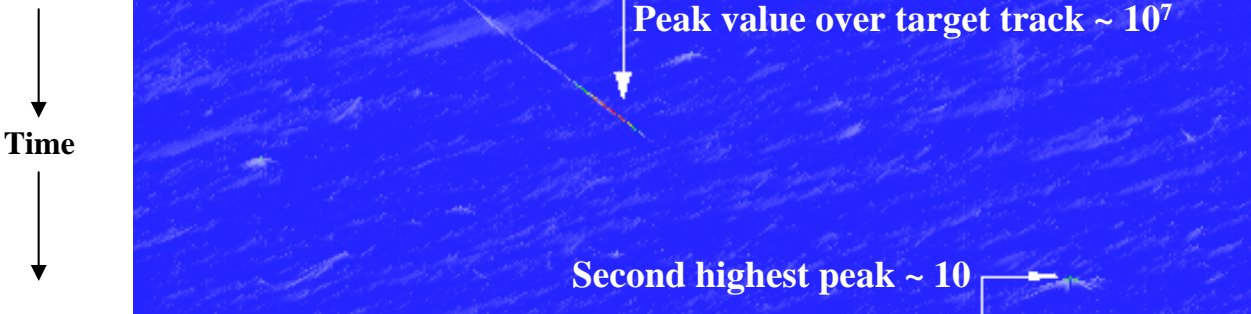
$$\Lambda^-(x_t) = \int P^T(x_t | x_{t-\Delta t}) \Lambda(x_{t-\Delta t}) dx_{t-\Delta t} \quad \Lambda(x_t) = \mathcal{L}(z_t | x_t) \Lambda^-(x_t)$$

Classical Likelihood Ratio Detection and Tracking (LRDT)

Radar return data –
Measurement likelihood ratio surfaces



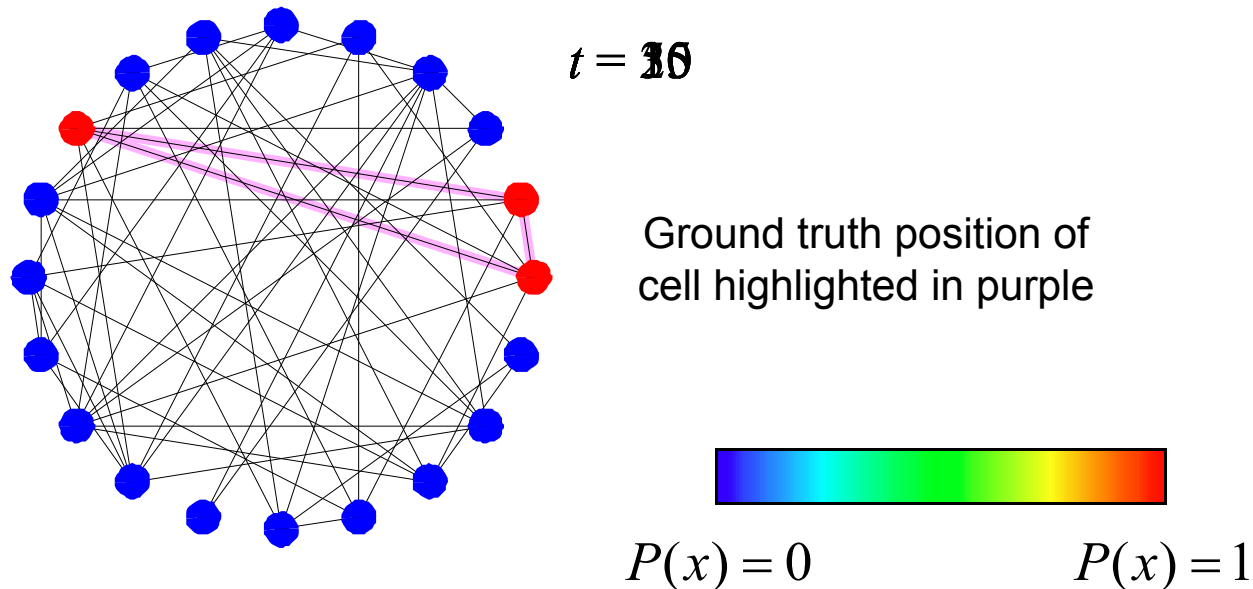
Cumulative likelihood ratio surface
(fusing data using motion model)



- Motion model describes movement of periscope
- Fusion over time smooths out random fluctuations from noise and clutter
- LR peaks accumulate on movement that fits the motion model (i.e., the periscope)
- LR peaks dissipate for structures which move in other ways (e.g., waves)

LRDT on Networks: a Simple Example

- State space: all 1140 possible triangles (i.e., “terrorist cells”)
- Measurement model:
 - Signal: the cell appears with probability 0.8
 - Noise: each possible edge appears independently with probability 0.3
- Motion model: cell swaps out a member with probability 0.1



Overview

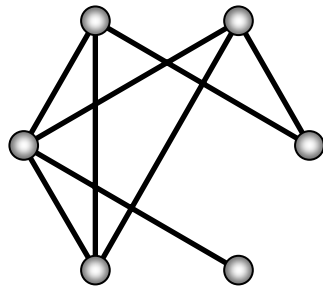
1. Graph-theoretic Underpinnings
 - Graph-theoretic analogues of noise and signal
2. Tracking Plans in Networks
 - Extension of LRDT to transactional domain
3. Hierarchical Hypothesis Management
 - Novel methodologies required to mitigate combinatorial explosion of state space
4. References

1. Graph-theoretic Underpinnings

- Erdős-Rényi Random Graph Model $G(n,p)$
 - Provides noise model for detection problem
 - Simplest, most tractable random graph model
- Inserted subgraph problem
 - Provides signal model for detection problem
- Likelihood ratio
 - Optimal decision statistic for detection of inserted subgraph
- Distribution of subgraph count
- Other random graph models?

Erdős-Rényi Random Graph Model $G(n,p)$

- The notation $G(n,p)$ denotes a random graph...
 - on n vertices
 - with each edge appearing independently with probability p



Instance of $G(n,p)$ for $n = 6, p = 0.5$

- Very simple noise process model
 - No correlation structure
 - Well-studied, but still yields difficult problems
 - First case to explore before moving on to more realistic network models (Random Collaboration, Geometric, Gaussian, etc.)

Inserted Subgraph Problem

- Evidence graph J
 - Background noise Erdős-Rényi random graph $G(n,p)$.
 - Target graph H may or may not be inserted somewhere.
- Binary decision problem: Is H present or not?
- Neyman–Pearson lemma: Likelihood ratio

$$\Lambda_H(J) = \frac{P(J \mid H \text{ present})}{P(J \mid H \text{ not present})}$$

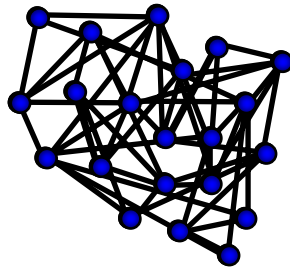
is the optimal decision statistic, i.e., yields highest probability of detection for given false alarm rate.

- **Theorem** [Mifflin, Boner]:

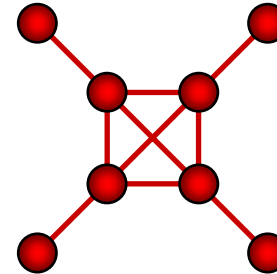
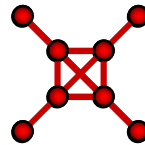
$$\Lambda_H(J) = \frac{X_H(J)}{\mathbb{E}[X_H]} = \frac{\# \text{ copies of } H \text{ in } J}{\# \text{ copies of } H \text{ expected just from noise}}$$

Inserted Subgraph Problem

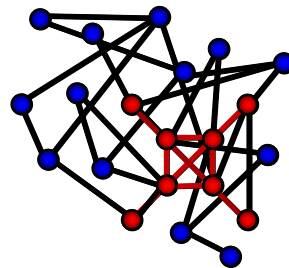
People, Companies,
Ports, Groups, Etc.



Target
Graph H



H could represent
four shipments of
precursor items to
four distinct, but
linked entities



Noise
Process

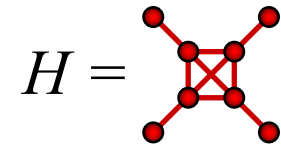
Signal + Noise
Process

Likelihood Ratio

$$\Lambda_H(\mathcal{J}) = \frac{P(\mathcal{J} \mid \text{dots} + \text{red graph})}{P(\mathcal{J} \mid \text{dots})}$$

optimal decision
statistic

Likelihood Ratio Calculations



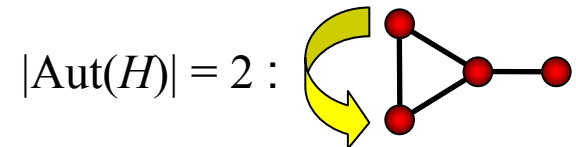
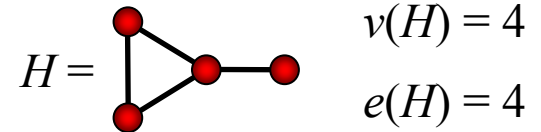
- **Example 1:** $n = 100, p = 0.07, X_H(J) = 200$
 - Are the 200 copies of H likely to have arisen by chance?
 - Answer:
$$\Lambda_H(J) = \frac{X_H(J)}{\mathbb{E}[X_H]} = \frac{200}{883.1} = 0.226 < 1 : \text{probably just noise}$$
- **Example 2:** $n = 1000, p = 0.007, X_H(J) = 2000$
 - Are the 2000 copies of H likely to have arisen by chance?
 - Answer:
$$\Lambda_H(J) = \frac{X_H(J)}{\mathbb{E}[X_H]} = \frac{2000}{11.4} = 174.8 \gg 1 : \text{probably contains target(s)}$$
- **Need information about distribution of X_H to set thresholds and establish performance boundaries!**
 - Expected value of X_H easy.
 - But that's all that's easy about it!

Expected Value of Subgraph Count X_H

- $\mathbb{E}[X_H]$: average # of subgraphs in an instance of $G(n,p)$

- Some preliminary notation

- $v(H)$ = number of vertices of H
- $e(H)$ = number of edges of H
- $|\text{Aut}(H)|$ = number of automorphisms of H



- Simple formula for $\mathbb{E}[X_H]$ (Erdős):

$$\mathbb{E}[X_H] = \underbrace{\binom{n}{v(H)}}_{\text{# of choices for vertex set of } H} \underbrace{\frac{v(H)!}{|\text{Aut}(H)|}}_{\text{# of arrangements of } H \text{ on this vertex set}} \underbrace{p^{e(H)}}_{\text{probability of all } e(H) \text{ edges of } H \text{ appearing}}$$

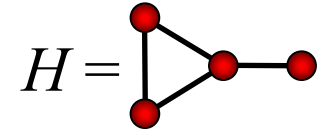
of choices for vertex set of H

of arrangements of H on this vertex set

probability of all $e(H)$ edges of H appearing

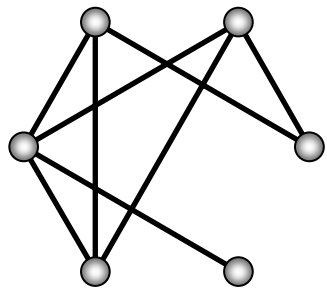
$$\begin{aligned} \mathbb{E}[X_H] &= \binom{n}{4} \frac{4!}{2} p^4 \\ &= 180 p^4 \\ &\text{for } n = 6 \end{aligned}$$

Distribution of Subgraph Count X_H

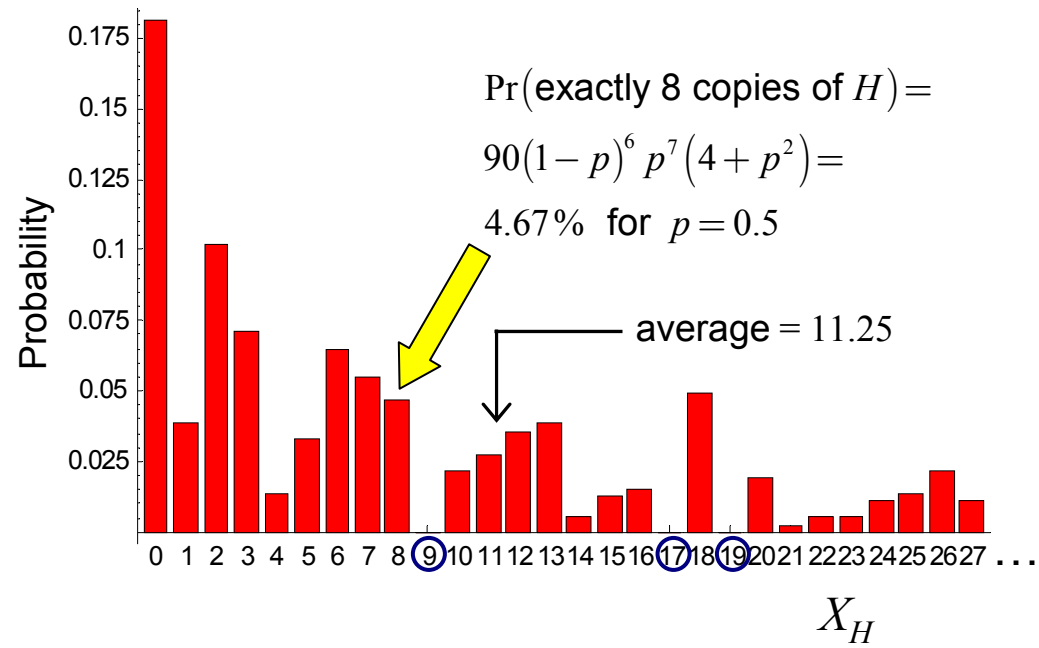
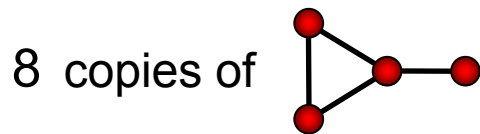


- Example: distribution of X_H for $n = 6, p = 0.5$
 - Minimum possible value = 0 (probability = 18.1%),
 - Maximum possible value = 180 (probability = 0.003%)
 - Mean given by Erdős formula:

$$\mathbb{E}[X_H] = 180p^4 = 11.25$$



Instance of $G(n, p)$ for $n = 6, p = 0.5$



- Variance:

$$\text{var}[X_H] = 180p^4(1-p)(1+13p+50p^2+128p^3) = (14.2)^2$$

Asymptotic Estimate of Variance

■ Theorem [Ferry]:

- Decompose H into core $cr(H)$ and rooted trees T_i
- Color each node i of core by isomorphism class of T_i
- Then

$$\frac{\text{var}[X_H]}{\mathbb{E}[X_H]} = \sum_{\pi \in G} \prod_{i \in V[cr(H)]} B(T_i, T_{\pi(i)}; d) + O(n^{-1})$$

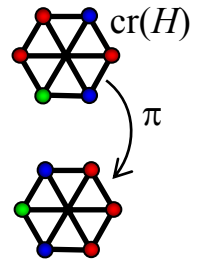
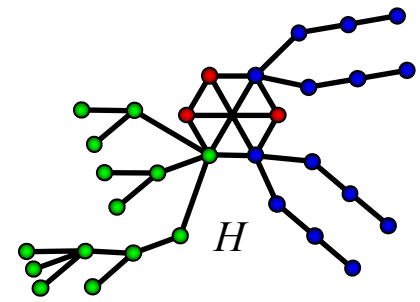
where

$$G = \text{Aut}[cr(H)] / \text{Aut}_{\chi}[cr(H)]$$

and B is computed by an exact recursive formula.

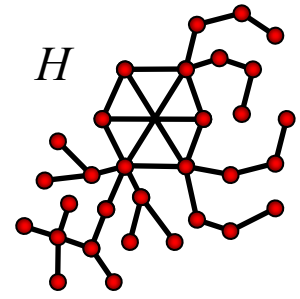
■ New result for study of distribution of X_H

- Bollobás, 1981 only applies to strictly balanced graphs.
- Ruciński, 1988 does not estimate variance.
- Janson, Łuczak, Ruciński give a much worse estimate.



$$\begin{aligned} & B(\text{green}, \text{blue}; d) B(\text{blue}, \text{red}; d) \times \\ & B(\text{red}, \text{red}; d) B(\text{blue}, \text{red}; d) \times \\ & B(\text{red}, \text{blue}; d) B(\text{red}, \text{green}; d) \end{aligned}$$

Asymptotic Estimate of Variance



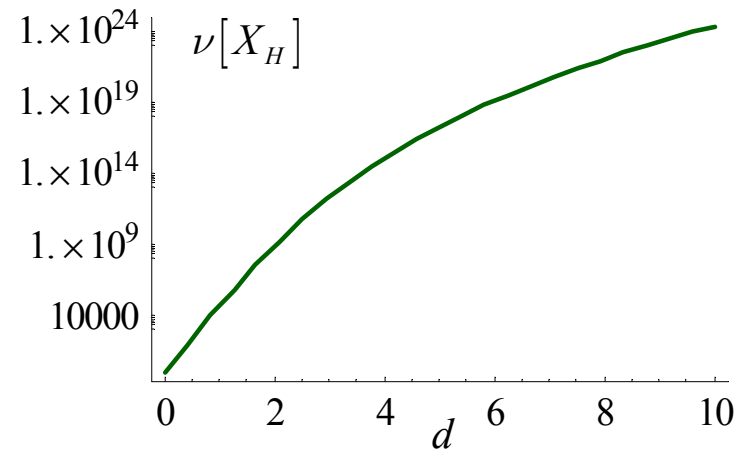
- With example H , formula yields

$$\frac{\text{var}[X_H]}{\mathbb{E}[X_H]} = \frac{1}{192} \left(192 + 2304d + 12960d^2 + 47008d^3 + 127120d^4 + 276544d^5 + 503280d^6 + 784840d^7 + 1066956d^8 + 1278644d^9 + 1361424d^{10} + 1296700d^{11} + 1110904d^{12} + 860672d^{13} + 607208d^{14} + 394332d^{15} + 239356d^{16} + 138208d^{17} + 76436d^{18} + 39914d^{19} + 19041d^{20} + 7875d^{21} + 2698d^{22} + 729d^{23} + 147d^{24} + 18d^{25} \right)$$

- For $n = 10^6$, mean degree $d = 10$

$$\begin{aligned} \mathbb{E}[X_H] &= \binom{n}{v(H)} \frac{v(H)!}{|\text{Aut}(H)|} \left(\frac{d}{n} \right)^{e(H)} \\ &= \binom{10^6}{31} \frac{31!}{768} \left(\frac{10}{10^6} \right)^{34} = 1.3 \times 10^{13} \end{aligned}$$

$$\text{var}[X_H] = \sqrt{\nu[X_H] \mathbb{E}[X_H]} = 5.4 \times 10^{18} \gg \mathbb{E}[X_H]$$

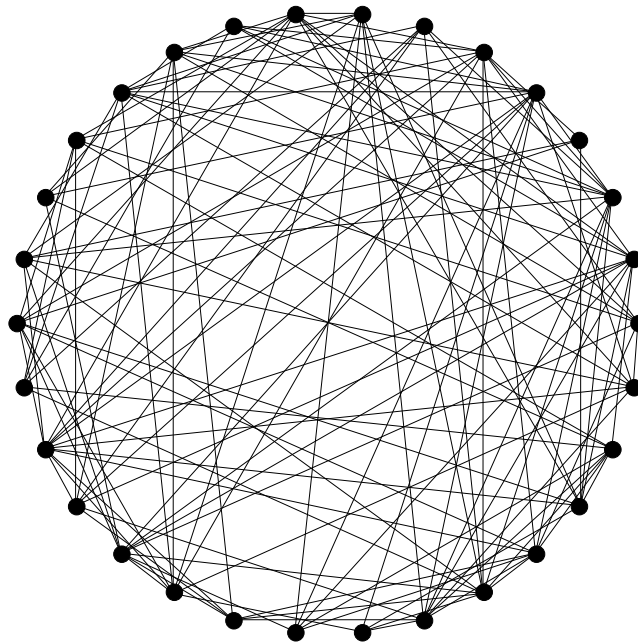


2. Tracking Plans in Networks

- Extension of LRDT to transactional domain
 - Noise model: Sequence of independent instances of $G(n,p)$
 - Signal model: Pattern of inserted subgraphs
- Susceptible to combinatorial explosion

Noise Model

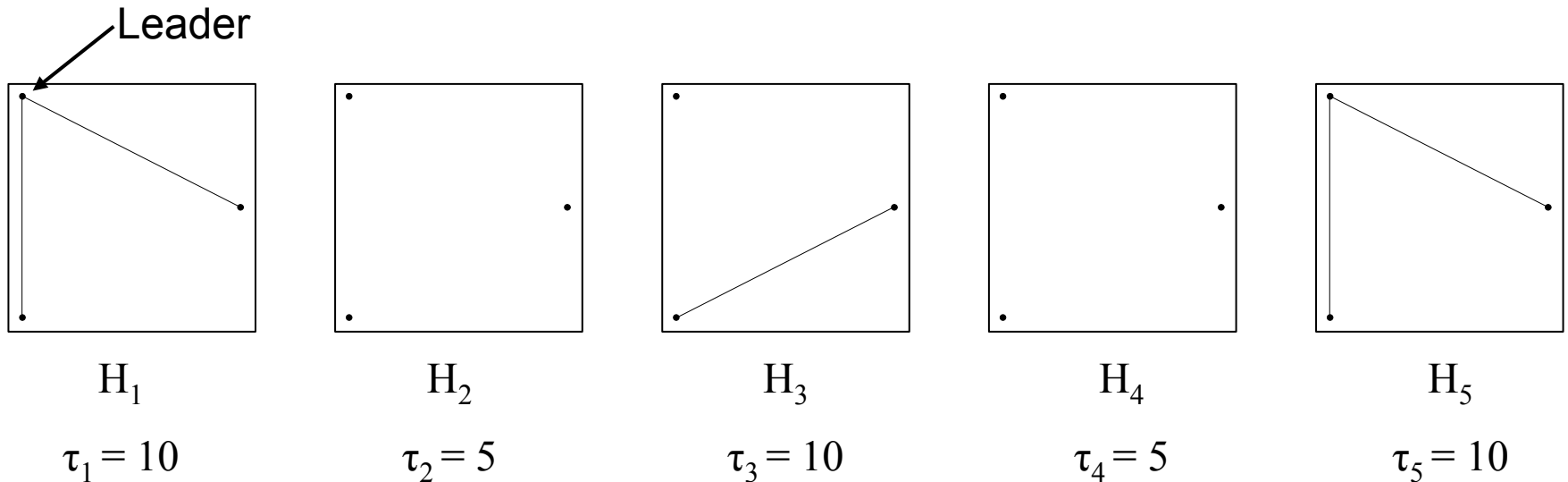
- Classic Erdős-Rényi random graph model $G(n, p)$
 - $n=30$
 - $p=0.3$
- Observe L instances J_1, J_2, \dots, J_L



$G(30, 0.3)$

Signal Model

- Insert a sequence of graphs H_1, H_2, \dots, H_m into some fixed, but unknown, location.
- The H_k s appear for τ_k time steps.
- Each edge has a probability p_V of being observed.
- Total Plan length: $T = 40$



State Space

- Must track location \tilde{H} and internal plan time τ .
- Possible time states:

$$1 \leq \tau \leq T, \tau = \alpha, \tau = \omega$$

α indicates the plan has not yet started.

ω indicates the plan has finished.

- State space:

$$\mathbf{X} = \left\{ (\tilde{H}, \tau) \right\}_{\tilde{H}, \tau} \cup \{\emptyset\}$$

\emptyset indicates no plan is present.

- Use a diffuse prior on the state space.

Size of the State Space

- There are $\binom{30}{3} \frac{3!}{2} = 12,180 \approx 10^4$

possible locations.

- Possible time states:

$$1 \leq \tau \leq 40, \tau = \alpha, \tau = \omega$$

There are 42 possible time states.

- State space consists of

$$(12,180)(42) + 1 = 511,561 \approx 5 \times 10^5$$

states.

Motion Model

- Advance the probability of each state as follows:

$$P^-((\tilde{H}, \tau), t) = \begin{cases} (1 - S(t))P((\tilde{H}, \alpha), t - 1) & \text{if } \tau = \alpha, \\ S(t)P((\tilde{H}, \alpha), t - 1) & \text{if } \tau = 1, \\ P((\tilde{H}, \tau - 1), t - 1) & \text{if } \tau = 2, \dots, T, \\ (P((\tilde{H}, T), t - 1) + P((\tilde{H}, \omega), t - 1)) & \text{if } \tau = \omega. \end{cases}$$

where

$$S(t) = \frac{1}{L - t + 1}$$

Measurement Model

- Let J be an evidence graph. The likelihood function is defined by

$$L(J|\emptyset) = p_{ER}^{e(J)} q_{ER}^{N-e(J)}.$$

and

$$L(J|(\tilde{H}, \tau)) = p_{ER}^{e(J \setminus \tilde{H}_k)} q_{ER}^{N-e(J \cup \tilde{H}_k)} p_*^{e(J \cap \tilde{H}_k)} q_*^{e(\tilde{H}_k \setminus J)}.$$

where $q = 1 - p$ and $q_* = q_{ER} q_V$.

Update Equation

- Update the probability distribution by

$$P(x, t) = \frac{1}{C} L(J|x) P^-(x, t)$$

where

$$C = \sum_{x \in \mathbf{X}} L(J|x) P^-(x, t)$$

- Note that x can be \emptyset or (\tilde{H}', τ') .

Example: Plan starts at time $t = 25$

Embedded movie removed

Summary: Tracking Plans in Networks

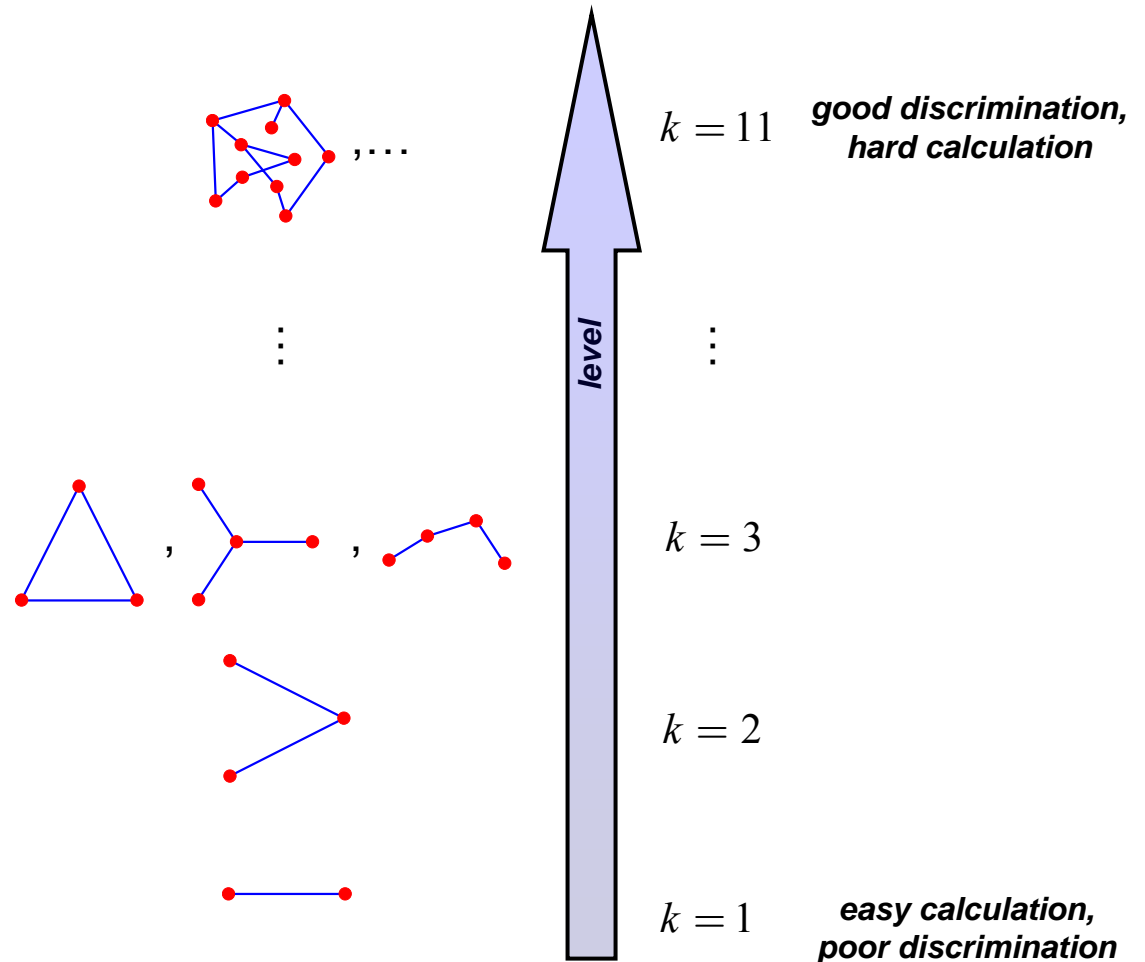
- Proof of concept: LRDT can be extended to transactional domain
- Can be generalized, e.g.:
 - May allow “tips” about possible subterfuge
 - May allow attributed nodes and links
 - Use more complicated network and plan models
 - May Include a clutter model
- Our best attempt to create a rule-based method to “score” nodes based on links and tips observed faltered
 - E.g. 6 malefactors identified, 1 of which is correct
 - Misled by tips
- Difficulty: number of states grows quickly

3. Hierarchical Hypothesis Management

- Number of hypotheses in previous example: 5×10^5
 - Numerically feasible to compute exactly
- For larger problems:
 - Number of hypotheses rapidly increases
 - Impossible to maintain all hypotheses
- Solution:
 - Group detailed hypotheses into successively coarser ones
 - Maintain probabilities on coarser hypotheses
 - High probability coarse hypotheses get resolved to finer levels
 - Low probability hypotheses perish
- Example:
 - Coarse hypothesis: “This edge is a member of the sought pattern”
 - Finer hypothesis: “This set of edges is a subset of the pattern”
 - Finest hypothesis: “This set of edges is the sought pattern”

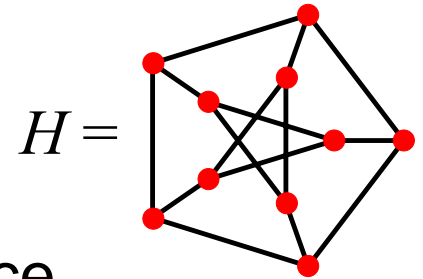
Feature Selection

- As features become larger...
 - they serve better to distinguish target from noise;
 - computational intensity increases.

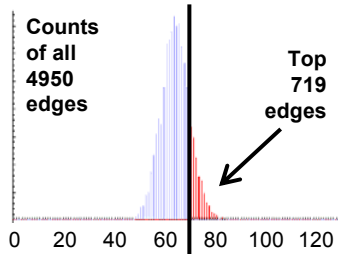


Example of HHM

- A larger problem: Find a given pattern H inserted 20% of the time In a fixed, unknown location (out of 5.2×10^{17} possible locations) In 125 instances of random noise on 100 nodes.
- Too hard for direct solution: in each instance, 160 trillion random copies of the pattern H obscure the real one.
- HHM algorithm recursively prunes hypothesis space until feasible to detect the target. final output is the configuration of the target pattern in the data.



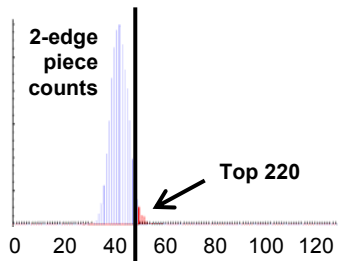
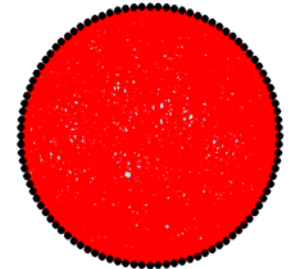
HMM in Action



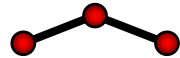
1-edge pieces



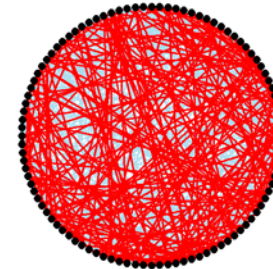
Pick top 719 edges out of 4950.



2-edge pieces



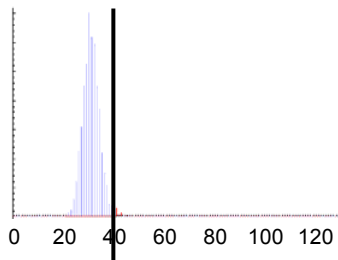
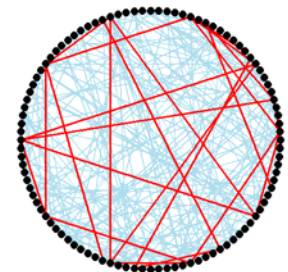
In these 719 are 10222 2-edge pieces.
Pick top 220 of these.



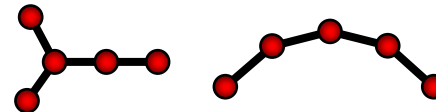
3-edge pieces



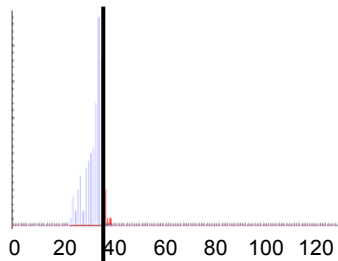
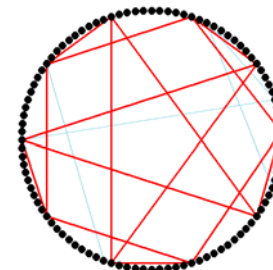
Among edges in top 220 2-edge pieces
are 2553 3-edge pieces.
Pick top 28 of these.



4-edge pieces



Among edges in top 28 3-edge pieces,
there are 145 4-edge pieces.
The edges in top 17 of these form the
sought pattern. Pattern found!



HHM in Action

Embedded movie removed

Summary: HHM

- LRDT approaches can have enormous state spaces
- In the classical domain, multigrid and particle methods have been developed to tame this problem
 - Both these rely on the existence of an underlying metric space
- In the intelligence domain, state spaces often lack a metric structure, so new technology is needed
- States can often be grouped into natural, user-defined hierarchies, which are then amenable to HHM
- Key research areas:
 - Optimal threshold setting, e.g. better than Monte Carlo
 - Feature selection, e.g. connected k -sets are more discriminating than arbitrary k -sets.

4. References

- J. Ferry and D. Lo, “Fusing Transactional Data to Detect Threat Patterns,” Proc. 9th International Conference on Information Fusion, Florence, Italy, July 2006.
- G. Godfrey, J. Cunningham and T. Tran, “A Bayesian, nonlinear particle filtering approach for tracking the state of terrorist operations,” Proceedings of the Military Applications Society Conference on Homeland Security in the 21st Century, Mystic CT, July 2006.
- T. Mifflin, C. Boner and G. Godfrey, “Detecting Terrorist Activities in the 21st Century: A theory of detection for transactional networks,” Emergent Information Technologies and Enabling Policies for Counter-Terrorism, eds. R. Popp and J. Yen, Wiley IEEE, June 2006.
- C. Boner, “Novel, Complementary Technologies for Detecting Threat Activities within Massive Amounts of Transactional Data,” Proceedings of the International Conference on Intelligence Analysis, Tysons Corner, May 2005.
- C. Boner, “Automated Detection of Terrorist Activities through Link Discovery within Massive Databases,” Proceedings of the AAAI Spring Symposium on AI Technologies for Homeland Security, Palo Alto, March 2005.
- T. Mifflin, C. Boner, G. Godfrey and J. Skokan, “A random graph model of terrorist transactions,” Proceedings of the IEEE Aerospace Conference, Big Sky, March 2004.