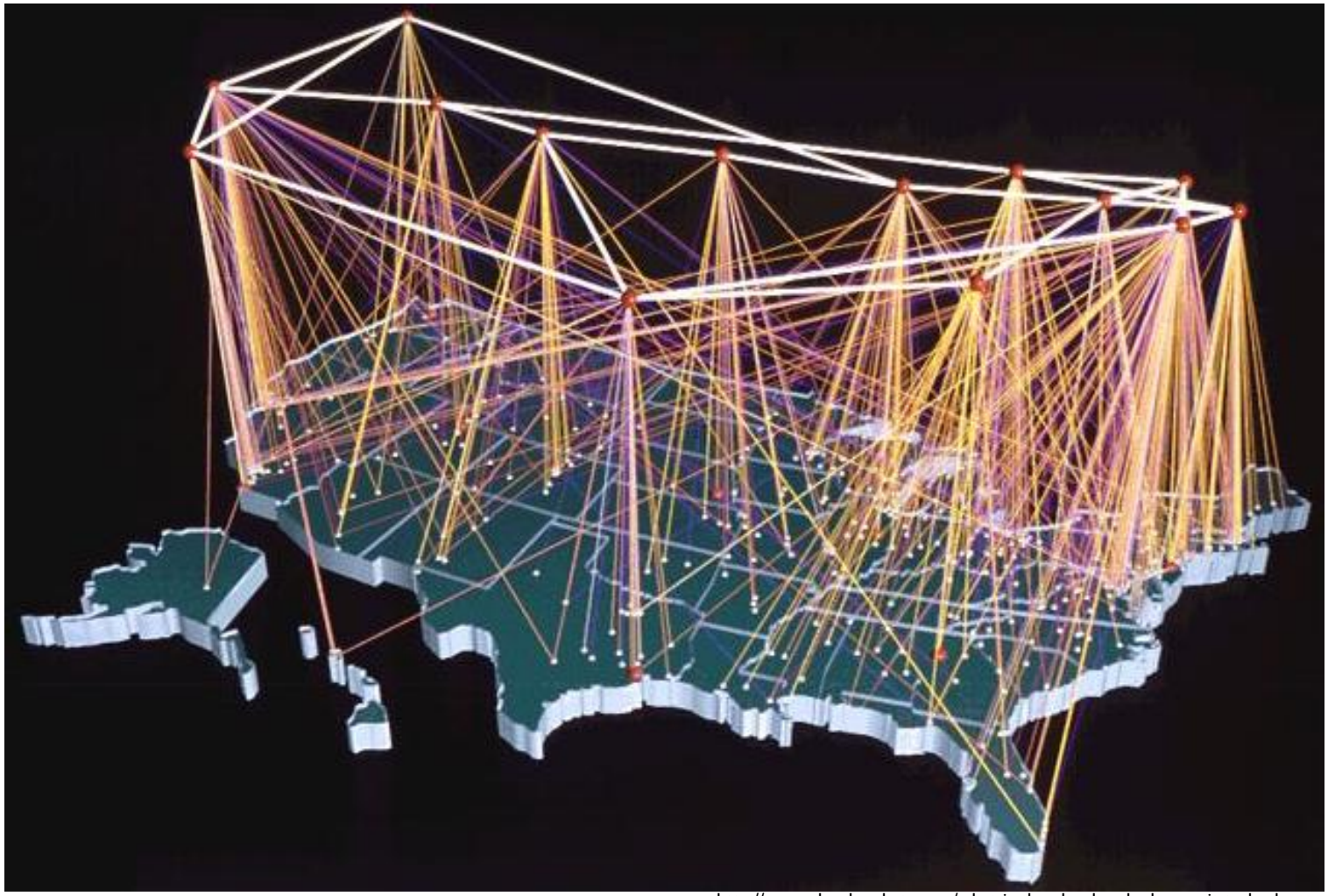
A Probabilistic Ontology for Large-Scale IP Geolocation

Kathryn Blackmond Laskey
Sudhanshu Chandekar
Bernd-Peter Paris

Volgenau School of Engineering
George Mason University

Tenth International Conference on Semantic Technology for Intelligence, Defense and Security

Large Scale, Discrete Geolocation



http://www.cloudpockets.com/what-is-the-cloud-and-why-use-it-to-backup

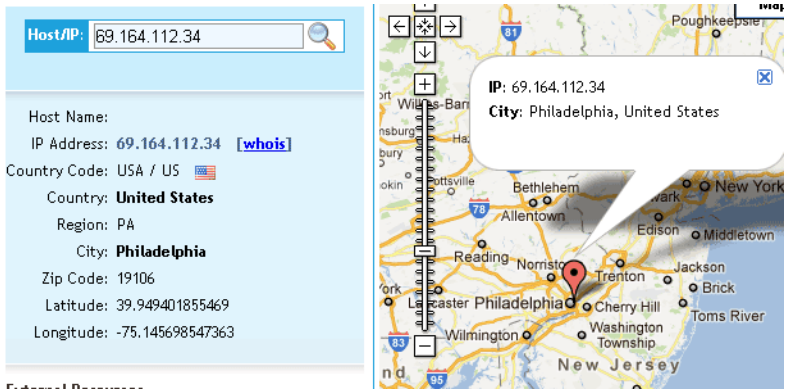
Problem: Given a set of IP addresses and a set of geographical regions, assign each address to a region

Problem Background

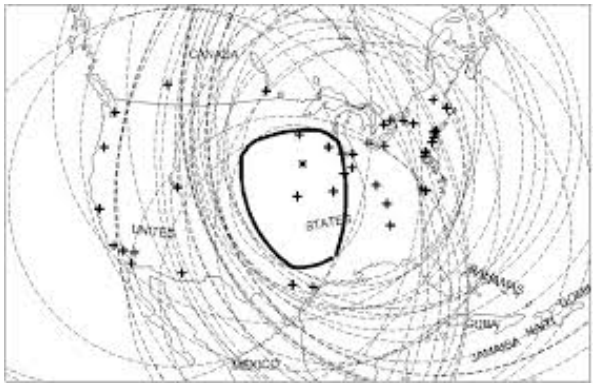
- Traditional geolocation techniques
  - Find precise location for a single address
  - Are commonly used for location-based services
- Large-scale, discrete geolocation
  - Many applications require less precise localization of many hosts
    - Cyber-situation awareness
    - Identifying source of coordinated attack
  - Traditional techniques do not scale

Data Sources

- Geolocation database
  - Used in location-based services
  - Focus on end-host location
  - Cannot effectively geolocate routers
- Hostname lookup
  - Can geolocate routers
  - Parse hostname for location clues
    - "0.ae1.br2.iad8.alternet" contains airport code IAD
- Delay measurement
  - Propagation delays depend on distance
  - Lack of scalability due to need for multiple landmarks and redundant measurements
- *All these data sources contain noise / errors*

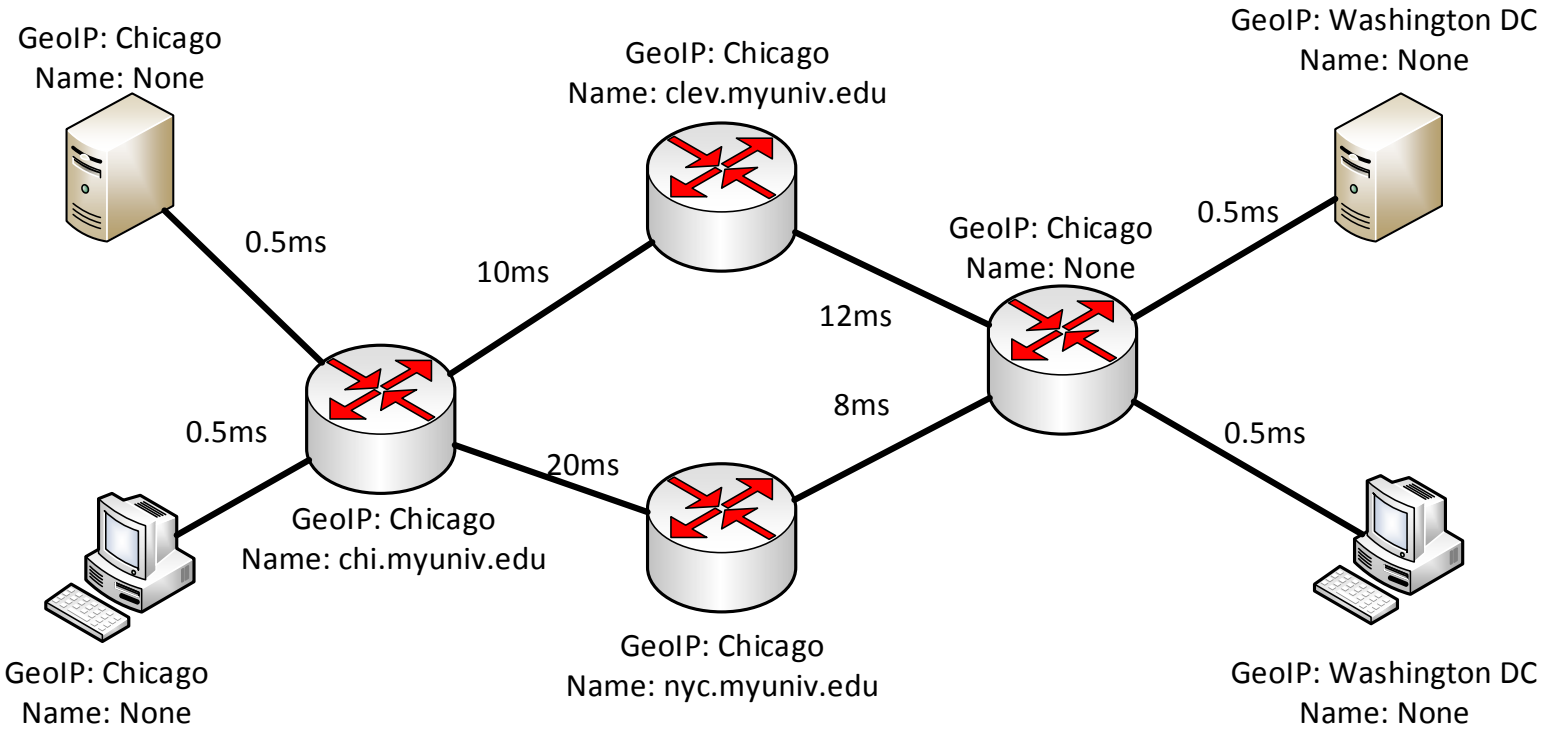


http://www.geoipview.com



http://fholzhauer.de/wp-content/uploads/2006/10/ip_geolocation_handout.pdf

Insight: Combine Data Sources

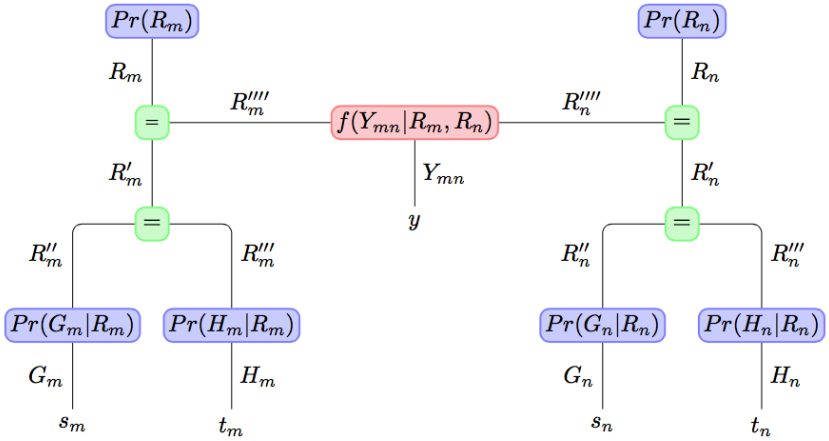


- Use node-local information from geolocation databases and hostname queries to localize single nodes
- Infer both topology and relative node separation from delay measurements

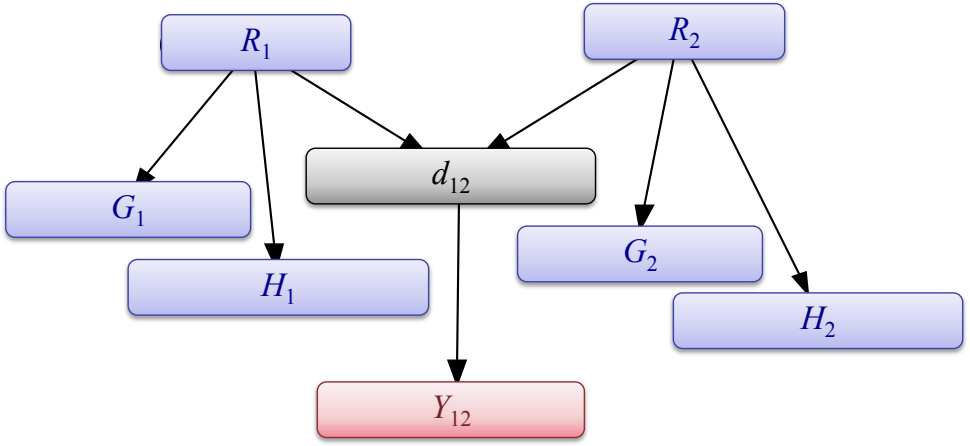
Data Sources

- **`GeoIP`** from MaxMind – freely available database for end host IP node geolocation
- **nslookup** tool – provides name given an IP address
  - Parse name to identify geographic clues
- **`DIMES`** topological database – provides host-to-host propagation delays using traceroute measurements

Graphical Probability Model for Geolocation in 2-Node Network



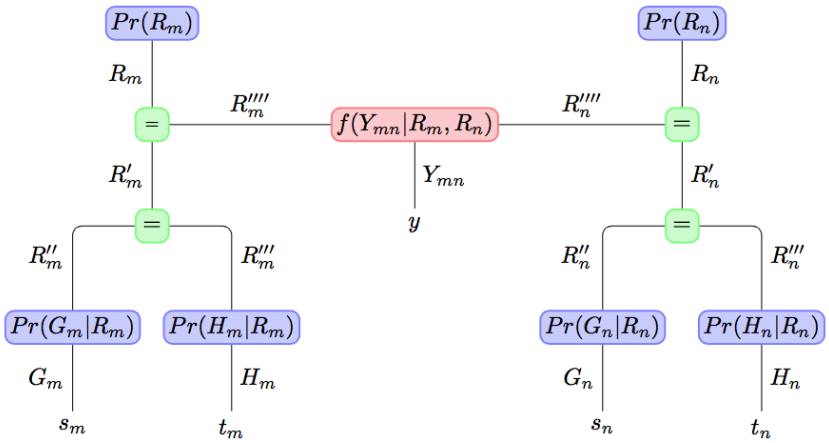
Factor graph model (Chandekar and Paris, 2015)



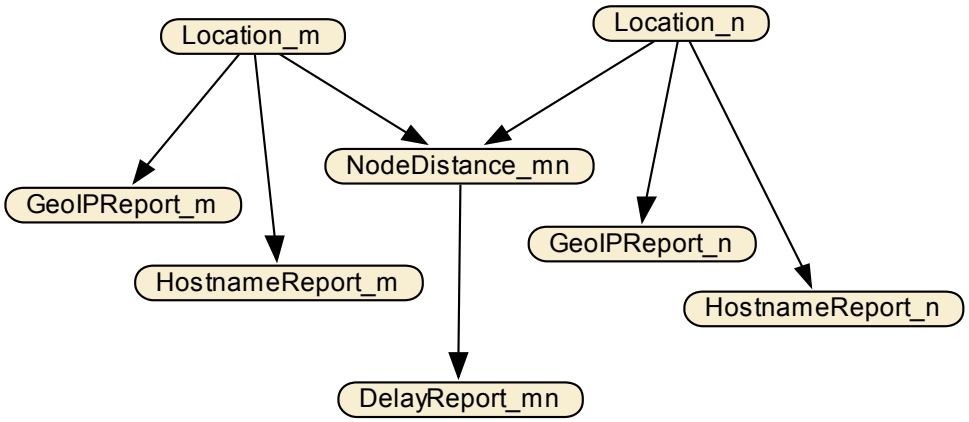
An equivalent Bayesian network

- Node locations (R_n) are uniformly distributed *a priori* among regions
- Hostname (H_n) reports are correct with probability β, and probability (1-β) is uniformly distributed over the incorrect regions
- Geolocation database reports (G_n) are correct with probability α, and probability (1-α) is uniformly distributed over the incorrect regions
- Delay measurements (Y_{mn}) have mixture of normal distributions with means depending linearly on distance (d_{mn}) between nodes

Graphical Probability Model for Geolocation in 2-Node Network



Factor graph model (Chandekar and Paris, 2015)

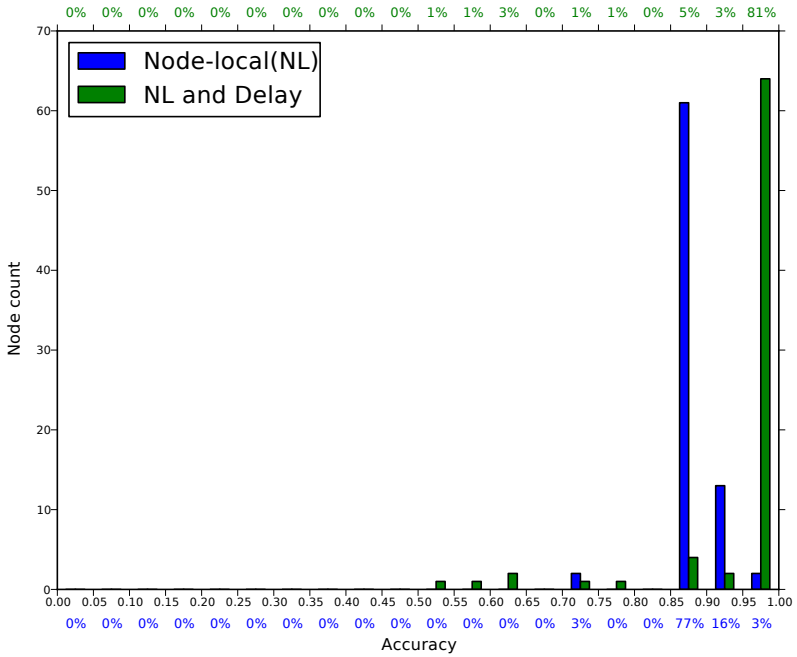


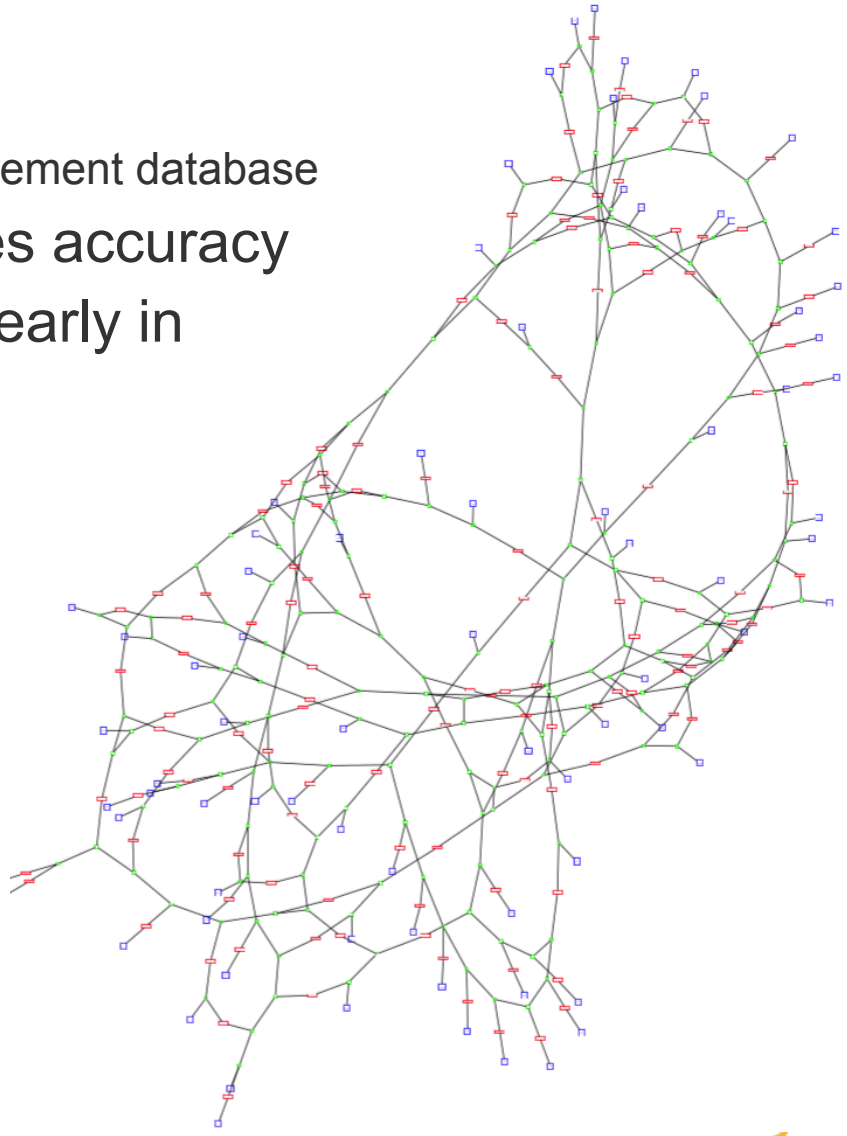
An equivalent Bayesian network

- **Node locations (R_n)** are uniformly distributed *a priori* among regions
- **Hostname reports (H_n)** are correct with probability β, and probability (1-β) is uniformly distributed over the incorrect regions
- **Geolocation database reports (G_n)** are correct with probability α, and probability (1-α) is uniformly distributed over the incorrect regions
- **Delay measurements (Y_{mn})** have mixture of normal distributions with means depending linearly on distance (d_{mn}) between nodes

Apply Model to Larger Network

- 79 nodes and 132 links
  - Links derived from DIMES delay measurement database
- Combining data sources improves accuracy
- Approximate inference scales linearly in network size





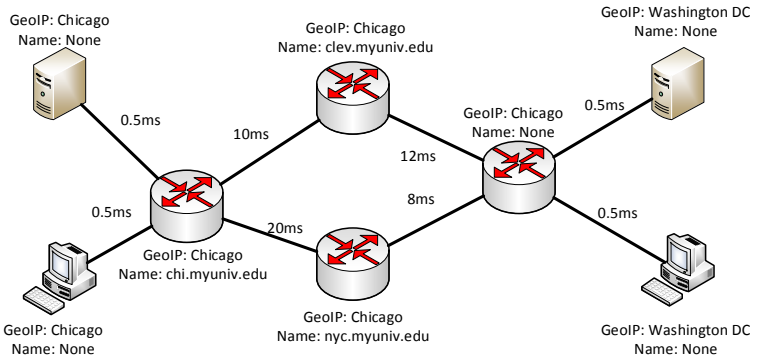
(Chandekar and Paris, 2015)

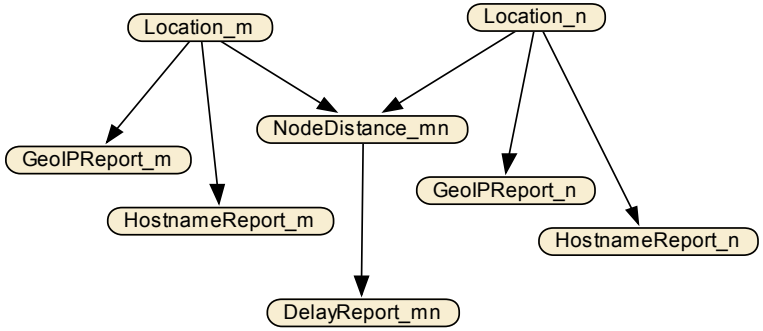
Semantic Models for IP Geolocation

- IP geolocation is usually one aspect of a larger capability
  - Identify source of a pattern of cyber attacks
  - Provide cyber situation awareness
  - Examine geographic patterns in internet usage
- Geolocation services need to interoperate smoothly with other elements of a cyber security tool suite
- Explicitly representing semantics supports interoperability and reuse

Types, Properties and Relationships

Entity	Property	Description
IPNode	Location	Region in which IP node is located
Region	RegionID	Unique identifier for a region
ProbePacket	StartingNode	Starting node for a link delay measurement
	EndingNode	Ending node for a link delay measurement
EvidenceItem	ReportedNode	IP node to which a database query or hostname lookup refers
	GeoIPReport	Region returned by database query on IP node
	HostnameReport	Region returned by hostname lookup on IP node
	ReportedProbe	Probe packet to which a link delay measurement refers
	DelayReport	Measured delay for a probe packet sent across a link

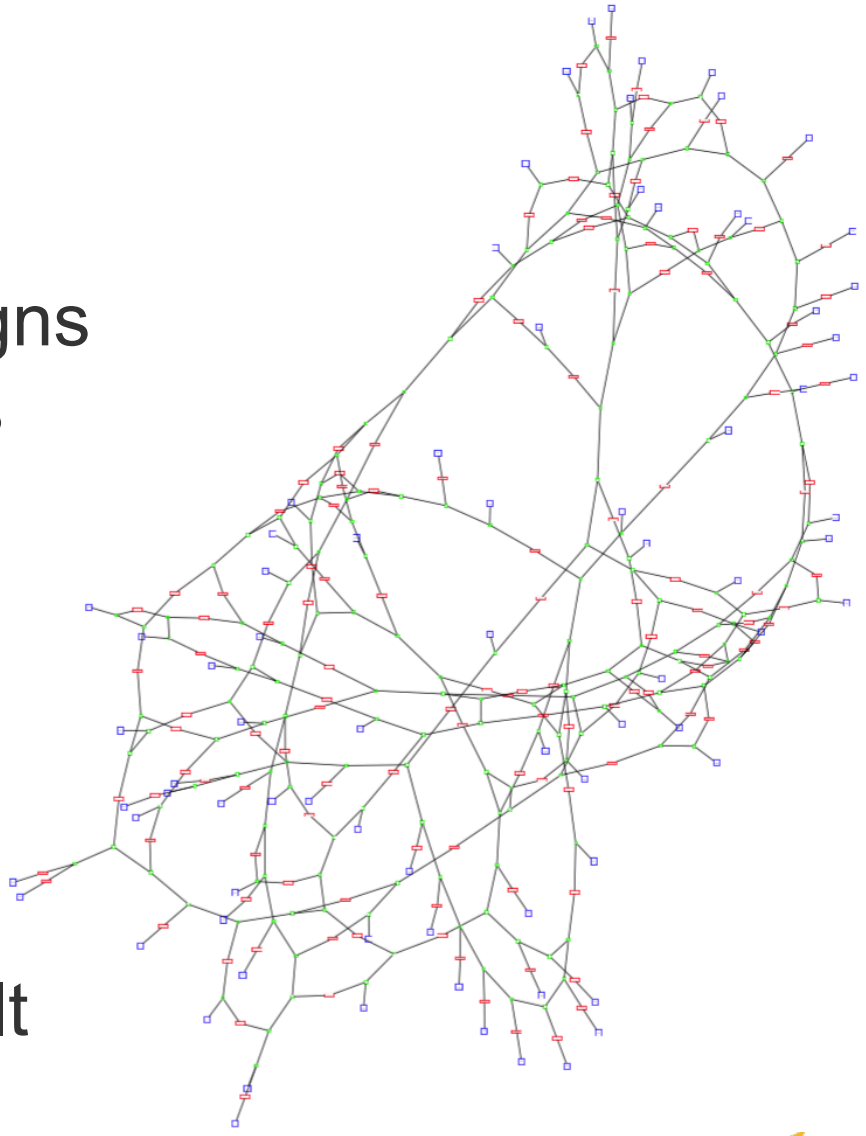




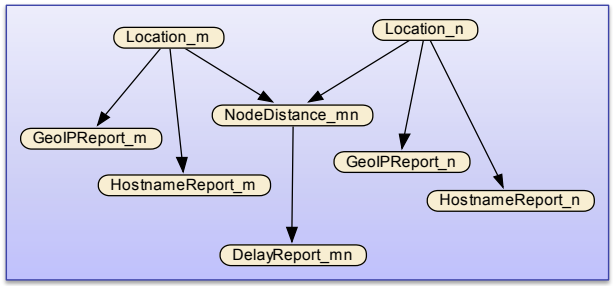
Relationship	Entities	Description
IsA	Thing, Type	Indicates that an entity is of the referenced type
NodeDistance	IPNode, IPNode	Distance between two IP nodes (real number)
RegionDistance	Region, Region	Distance between two regions (real number)

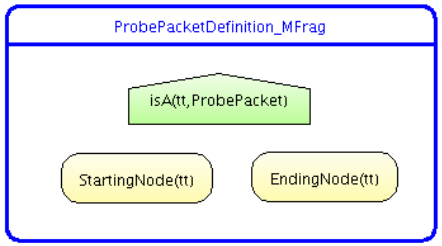
Representing Uncertainty

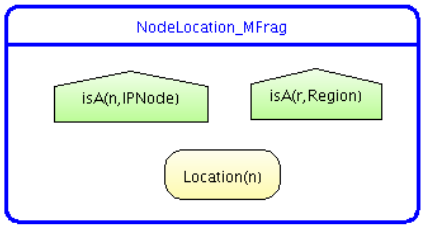
- Combining data sources reduces uncertainty
- Geolocation model assigns probabilities to attributes and relationships
- A probabilistic ontology can represent this uncertainty
- Semantics are explicit, not buried in custom-built code

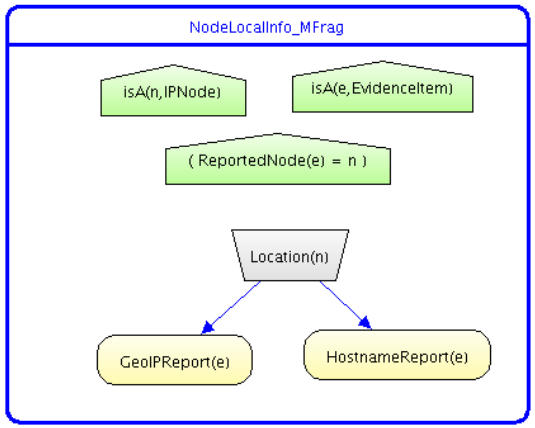


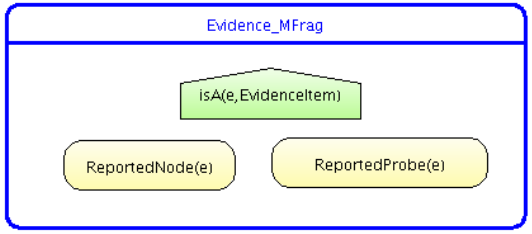
IP Geolocation PO

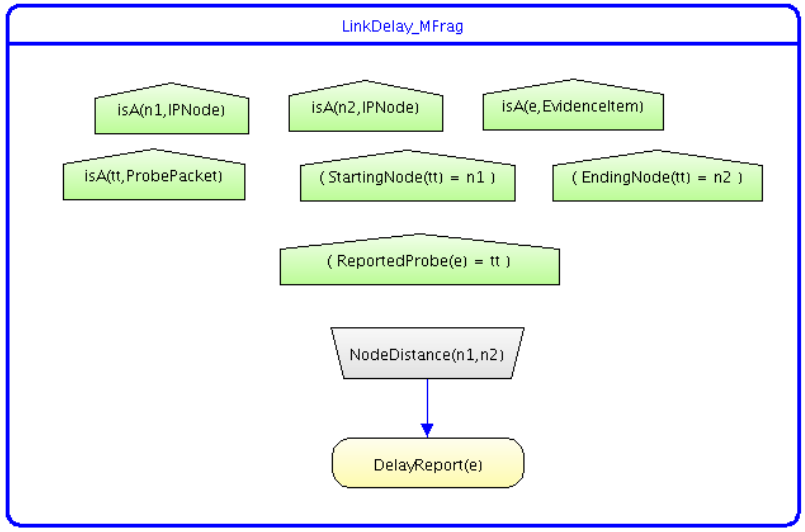


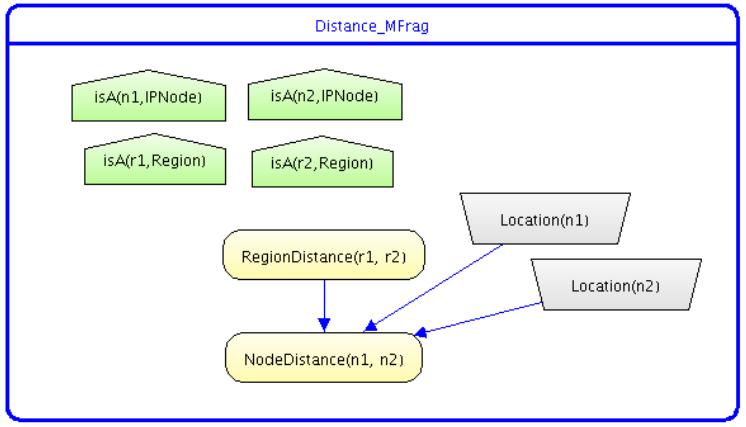






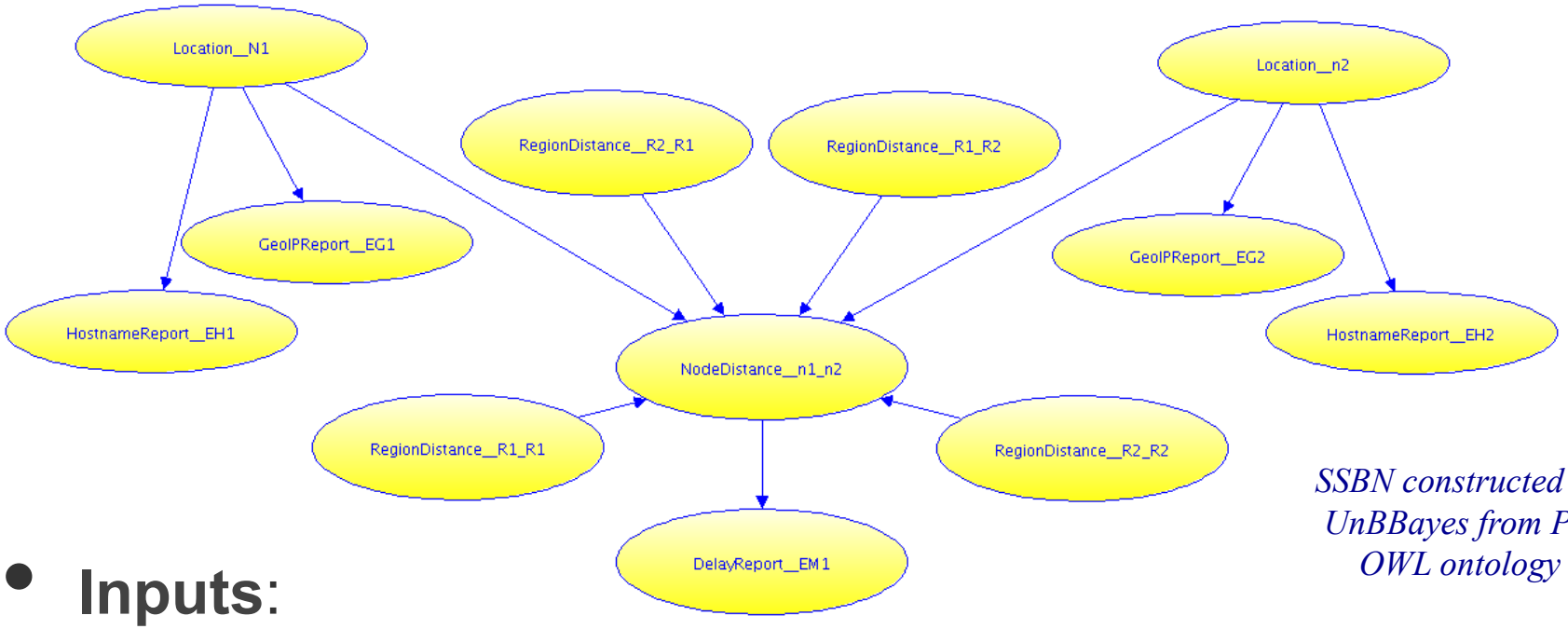






Pr-OWL ontology implemented in UnBBayes

Situation-Specific Bayesian Network
(for 2-Node Geolocation)



SSBN constructed by UnBBayes from Pr-OWL ontology

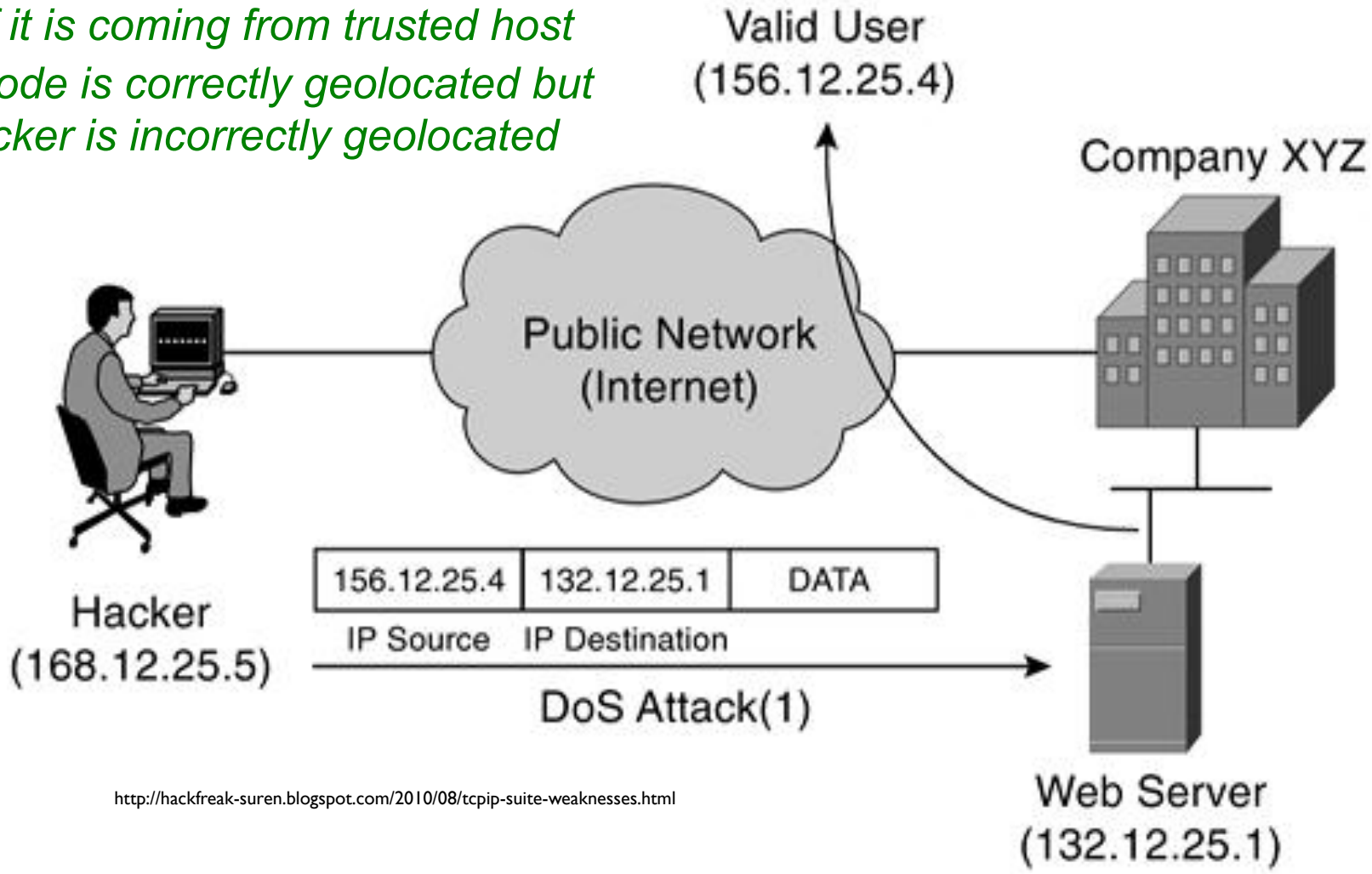
- **Inputs**:
  - A set of regions with inter-region distances
  - GeoIP, hostname and delay reports for a set of IP nodes
  - GeoIP probabilistic ontology
- **Output**: Bayesian network representing probable locations of all IP nodes

Benefits of Probabilistic Ontology

- Represent domain semantics with uncertainty
  - Location of node
  - GeoIP, hostname and delay reports
- Integrate logical and probabilistic reasoning in mathematically well-founded way
- Separate knowledge representation from fusion algorithm implementation
  - Free modeler from having to write custom algorithm
  - Easily extend algorithm innovations to new problem domains
- Extend / embed geospatial PO into larger cybersecurity PO

Extension Example: IP Spoofing

- *Attacker modifies packet to appear as if it is coming from trusted host*
- *IP node is correctly geolocated but attacker is incorrectly geolocated*



http://hackfreak-suren.blogspot.com/2010/08/tcpip-suite-weaknesses.html

Incorporating Address Spoofing

- Spoofing can often be detected
  - E.g., packet arrives along a link that is incompatible with alleged source address
  - E.g., outgoing packet with external source IP address
- The PO could be extended to detect spoofed IP addresses
  - Add `user` class
  - Users may be valid or spoofers
  - Spoofers need not be co-located with source IP address of packets they send
- Geolocating spoofers
  - Individual messages can in principle be back tracked but this is very difficult in practice
  - Attacks usually must have non-spoofed elementsthat can be used to geolocate attack source

Future Work

- Apply SSBN construction algorithm to 79-node network and compare with custom-built factor graph model
  - Accuracy and computation time
- Compare with best-of-breed approximate inference algorithms
- Extend to internet-scale geolocation (millions of nodes)
- Investigate integrating IP geolocation capability with other cybersecurity capabilities

C4I Center

GEORGE MASON UNIVERSITY

Thank you for your Attention!