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

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

- Node locations \((R_n)\) are uniformly distributed \textit{a priori} among regions
- Hostname \((H_n)\) reports are correct with probability \(\beta\), and probability \((1-\beta)\) is uniformly distributed over the incorrect regions
- Geolocation database reports \((G_n)\) are correct with probability \(\alpha\), and probability \((1-\alpha)\) 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

Factor graph model (Chandekar and Paris, 2015)

An equivalent Bayesian network
Graphical Probability Model for Geolocation in 2-Node Network

- **Node locations** ($R_n$) are uniformly distributed a priori among regions
- **Hostname reports** ($H_n$) are correct with probability $\beta$, and probability $(1-\beta)$ is uniformly distributed over the incorrect regions
- **Geolocation database reports** ($G_n$) are correct with probability $\alpha$, and probability $(1-\alpha)$ 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

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

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Entities</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IsA</td>
<td>Thing, Type</td>
<td>Indicates that an entity is of the referenced type</td>
</tr>
<tr>
<td>NodeDistance</td>
<td>IPNode, IPNode</td>
<td>Distance between two IP nodes (real number)</td>
</tr>
<tr>
<td>RegionDistance</td>
<td>Region, Region</td>
<td>Distance between two regions (real number)</td>
</tr>
</tbody>
</table>

---

![Diagram](image-url)
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)

- **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

SSBN constructed by UnBBayes from Pr-OWL ontology
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

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 elements that 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
Thank you for your Attention!