A Probabilistic Ontology for Large-Scale IP Geolocation

Kathryn Blackmond Laskey
Department of Systems Engineering and Operations Research
George Mason University
Fairfax, VA 22030
Email: klaskey@gmu.edu

Sudhanshu Chandekar and Bernd-Peter Paris
Department of Electrical and Computer Engineering
George Mason University
Fairfax VA 22030
Email: [schandek, pparis]@gmu.edu

Abstract—**Mapping IP addresses to physical locations is important for a host of cyber security applications. Examples include identifying the origin of cyber attacks, protecting against fraud in internet commerce, screening emails for phishing, and enforcing restrictions on commerce with sanctioned countries. Simultaneous geolocation of large numbers of IP hosts is needed for cyber situation awareness. Explicit formal representation of the geospatial aspects of the cyber domain is necessary for interoperation with other cyber security capabilities. Formally representing the uncertainty inherent in geolocation supports increased accuracy via information fusion, as well as integration of geospatial inference with inference about other aspects of the cyber landscape. This paper presents a probabilistic ontology (PO) for IP geolocation. The geolocation PO is represented in the PR-OWL language, which allows an OWL ontology to be augmented with information to support uncertainty management. We show how the PR-OWL ontology supports automated construction of a Bayesian network for simultaneously geolocating a large number of IP hosts. The ultimate aim is to integrate our probabilistic ontology into a comprehensive cyber security probabilistic ontology to support cyber situation awareness, predictive modeling, and response strategy definition.**

I. Introduction

Recognition is growing of the need to establish a common vocabulary for and a shared understanding of the cyber security domain (e.g., [1]). Explicit, formal representation of entity types, properties and relationships is a key means to this end ([2], [3], [4]). Among the advantages of such a cyber domain ontology include increasing interoperability of cyber security tools and methods, improving tools to support situation awareness among cyber security operators, and enhancing information sharing among domain experts (c.f., [5]). Anticipating, diagnosing and responding to increasingly sophisticated cyber threats requires drawing on and fusing information from diverse sources. Automated fusion of hard and soft, structured and unstructured information requires semantic as well as syntactic interoperability among information providers and consumers. Ontologies are a key enabler of semantic interoperability.

The cyber security domain is fraught with uncertainty. Support for uncertainty management is a key requirement for cyber situation awareness and decision support tools. Probabilistic ontologies augment traditional ontologies with the ability to represent uncertainty associated with properties of and relationships among domain entities, supporting semantically aware automated uncertainty management [6].

This paper presents a case study of the use of a probabilistic ontology to represent and reason about a key problem in the cyber security domain, mapping IP addresses to physical locations. Example applications of IP geolocation include identifying the origin of cyber attacks, protecting against fraud in internet commerce, screening emails for phishing, and enforcing restrictions on commerce with sanctioned countries. Most IP geolocation methods focus on identifying the location of a single IP host. To support cyber situation awareness, a useful capability is simultaneous geolocation of a large number of hosts, with a reduced requirement for geographic resolution.

As an essential component of an overall cyber security strategy, geolocation services need to interoperate smoothly with other elements of a cyber security tooklit. For this purpose, an ontology of the geospatial aspects of the cyber domain can form a useful module in a cyber domain ontology. Available information for IP geolocation is fraught with uncertainty. Representing the uncertainty inherent in geolocation can support more accurate geolocation through information fusion, as well as integration of geospatial inference with inference about other aspects of the cyber landscape.

This paper presents a probabilistic ontology (PO) for IP geolocation and describes its application to the simultaneously IP node geolocation problem. The geolocation PO is represented in the PR-OWL language, which provides constructs for augmenting an OWL ontology with information to support uncertainty management. We show how the PR-OWL ontology supports automated construction of a Bayesian network for simultaneously geolocating a large number of IP hosts. The ultimate aim is to integrate our probabilistic ontology into a comprehensive cyber security probabilistic ontology to support cyber situation awareness, predictive modeling, and response strategy definition.

The paper is organized as follows. Section II gives a brief overview of previous research on IP node geolocation. Section III presents a factor graph model [7] for simultaneous IP node geolocalization that forms the basis for our PO. Section IV makes the case for explicitly representing the semantics of the model, presents a probabilistic ontology for IP node localization, and shows how the probabilistic ontology can be queried using a generic reasoner to construct a geolocation model that is formally equivalent to the model of [7]. Section V presents a summary and discussion.

II. Background: IP Geolocation

Although most work on IP geolocation focuses on identifying the physical location of a single IP host, the problem of simultaneous geolocation of many IP hosts is beginning to receive attention ([7], [8]). The challenge for large-scale IP geolocation is three-fold. The first is to expand the scope of IP geolocation by taking into account not only hosts at the network edge but also hosts in the network core, such as routers. The second is to achieve scalability to large numbers of IP hosts, which requires the ability to simultaneously infer the location of many hosts. The final challenge is to improve geolocation accuracy while not sacrificing the first two objectives. To tackle these challenges, we introduce a model that uses Bayesian inference to fuse information from multiple sources to simultaneously geolocate a set of IP addresses to within a discrete set of geographic regions. The simultaneous geolocation model underlying our IP geolocation PO was presented in [7], and is reviewed briefly in this section.

By itself, an IP address provides no information about a host's location. Therefore, information from external sources is required to map an IP address to a physical location. Available information comes from different sources, each subject to uncertainty. Information sources for geolocation can be classified into three broad categories: database-based, name-based and measurement-based [9].

Geolocation databases [10] contain mappings between IP addresses and locations. These providers tend to focus on geolocating end-hosts, and consequently tend to be unreliable in geolocating routers. Moreover, it has been observed that databases tend to geolocate blocks of IP addresses to the location where they were initially registered — often the business address of the network provider. As a result, some geographically distributed blocks of IP addresses that may be geolocated to the same location. Location information about devices in the core of the network, such as routers, can often be inferred from the names assigned to them.

Name-based geolocation [11] uses information embedded in a hostname, such as an airport code or a city abbreviation, to infer the location of the host. For example, the hostname `ip68-100-3-241.dc.dc.cox.net` indicates a device located in Washington D.C. When available, hostname information tends to be fairly reliable, but it is not always available.

Measurement-based geolocation [12] uses network information such as delay and topology to estimate the location of nodes. When location of and connectivity to "landmark" hosts is available, measurement data can be used to infer the location of other nodes in the network. However, such techniques depend not only upon active landmarks that conduct delay measurements among themselves and the target but also on passive landmarks that are used for approximating the target's location. Also, due to factors discussed later, some delay measurements may be biased significantly. As a result, delay-based geolocation errors may be large, sometimes on the order of several hundred kilometers. The size of the error has been shown [13] to be correlated with the number of distributed landmarks and with the number of probes between landmarks and the unknown target. The dependence on many distributed hosts with known locations, coupled with the focus on pinpointing the location of individual target hosts, renders such techniques impractical for geolocating large numbers of IP hosts.

III. A Model for IP Geolocation

From the discussion above, it is clear that geolocating a large number of hosts requires coping with missing and/or imperfect information. Fusing information from the geolocation database and the location hints obtained from hostnames admit information about mutually exclusive sets of hosts, thereby increasing the number of hosts that can be geolocated. However, this does not guarantee improvement in accuracy due to the aforementioned uncertainties in the respective information sources. Because past studies [14], [15] have shown a strong relationship between measured delay and physical distance, we incorporate evidence about link delay into our model as a means to improve accuracy.

Our link delay measurements are taken from the DIMES database [16], [17], constructed as part of an ongoing, distributed, open-source project to map the structure and topology of the Internet. The DIMES database contains a set of `traceroute` measurements. Each `traceroute` measurement includes the measured round trip time (RTT) from a source node along a path toward a destination node, along with the IP addresses of the intermediate nodes along the path. Host-to-host or link delays can be inferred by subtracting RTTs for consecutive hosts. In addition, path information can be used to infer the network topology.

Our model assumes that IP nodes are to be geolocated into a set of M discrete disjoint geographic regions. A joint probability distribution is defined over the random variables defined in Table I. The random variable R_n denotes the region in which IP node n is located; G_n represents the result of a geolocation database query for IP node n; H_n represents the result of a hostname lookup for IP node n. The random variable Y_{mn} represents a measurement of relative host-to-host propagation delay between host m and host n.

TABLE I
Random Variable Definitions

Random Variable	**Definition**
R_n	Region in which node n is located
G_n	Region returned by geolocation database query for node n
H_n	Region returned by hostname lookup for node n
Y_{mn}	Delay measurement for signal transit between nodes m and n

A joint probability distribution for the random variables $\{R_n, G_n, H_n, Y_{mn}\}$ is defined in factored form as follows.

In the absence of prior information on the distribution of hosts across regions, independent uniform distributions are assumed for the node locations R_n. That is, for each region r,

$$\Pr(R_n = r) = \frac{1}{M}. \tag{1}$$

Of course, if information were available about the relative density of nodes in different regions, it could be encoded as an informative prior distribution on the locations R_n.

Conditional on the region in which n is located, the geolocation database and hostname evidence are independent with distributions given by:

$$\Pr(G_n = s|R_n = r) = \begin{cases} \alpha & r = s \\ \frac{1-\alpha}{M-1} & r \neq s \end{cases} \tag{2}$$

and

$$\Pr(H_n = t|R_n = r) = \begin{cases} \beta & t = r \\ \frac{1-\beta}{M-1} & t \neq r \end{cases} \tag{3}$$

That is, the model assumes there is a probability α that the geolocation database query returns the correct region, with the remaining probability distributed uniformly among the remaining regions. Similarly, there is a probability β that the hostname lookup returns the correct region, with the remaining probability distributed uniformly among the remaining regions. Again, this simple model could be modified to incorporate additional information if available.

Finally, the distribution of a link delay measurement depends on the distance between the starting and ending point of the link. A measurement probe packet traveling from a source to a destination may encounter other delays at each node. While propagation delay is directly proportional to distance, the true linear relation is distorted by the presence of other delays, including queueing, transmission and nodal processing delay. For Y_{mn} measuring link delay between nodes m and n, let D_{mn} denote the distance between the regions R_m and R_n where nodes m and n are located. Chandekar and Paris [7] considered a normal distribution for Y_{mn} given R_m and R_n:

$$\begin{aligned} f(Y_{mn} = y|R_m = r, R_n = s) &= \\ f(Y_{mn} = y|D_{mn} = d) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-ad-b)^2}{2\sigma^2}} \end{aligned} \tag{4}$$

with mean $ad + b$ a linear function of the distance between the starting and ending point, and standard deviation σ. They also noted that delay measurements may be biased due to the fact that some internet routers delay ICMP (Internet Control Message Protocol) replies. Thus, they also considered a Gaussian mixture model with different components for zero, positive and negative bias.

The distributions (1 - 4) are combined into a factored representation for the joint distribution of $\{R_n, G_n, H_n, Y_{mn}\}$. The joint probability mass / density function is:

$$\prod_{n \in \mathcal{N}} \Pr(R_n) \Pr(G_n|R_n) \Pr(H_n|R_n) \prod_{(m,n) \in \mathcal{L}} f(Y_{mn}|R_m, R_n) \tag{5}$$

where $\mathcal{N}$ is the set of IP nodes to be localized and $\mathcal{L}$ is the set of links, or pairs of nodes connected by signal transmission measurements.

Fig. 1, adapted from [7], depicts a factor graph for the joint distribution (5) when there are two nodes connected by a single link. A factor graph is a graphical probability model for a joint distribution represented in factored form [18]. The figure shows a Forney-style factor graph [19], [20], in which nodes are labeled by factors of the joint distribution and edges connect pairs of factors that share a random variable. Edges are labeled by the random variable shared between the factors at either end of the edge. If a random variable is shared by more than two factors, equality constraint nodes are inserted into the graph to "clone" random variables so each random variable is shared by no more than two factors. Evidence is shown as labels at the end of edges extending from the random variables whose values are observed. For example, evidence that G_m has value s_m is depicted at the terminus of an edge extending from the factor $\Pr(G_m|R_m)$.

From Fig. 1 it can be observed that the underlying physical topology determines the connectivity between random variables in the factor graph. The graph shows a cluster for each of the two hosts. Each of these clusters contains a factor for the prior distribution over the node's location, as well as a factor for evidence from the geolocation database query and a factor for evidence from the hostname lookup. These evidence items local to each host are henceforth referred to as node-local evidence. Fig. 1 also contains a link between the two clusters, which is labeled by a factor representing the delay measurement.

Extending this model to an arbitrary network results in a factor graph containing a node-local evidence cluster for each host in the network and an edge connecting any two clusters for which there is a link delay measurement. Thus, the factor graph structure mirrors the topology of the physical network. Each node-local evidence cluster corresponds to a physical IP node and the delay evidence edge corresponds to a connection (IP link) between nodes.

This mapping between network topology and factor graph allows for the systematic and simultaneous geolocation of a set of interconnected nodes using the joint probability distribution (5). This can be achieved by finding the joint posterior distributions of the node regions $\{R_n\}_{1 \leq n \leq M}$ conditional on database, hostname and delay evidence. For general network topologies, solving for the joint posterior distribution is intractable. However, the well-known sum-product algorithm [18] can be applied to estimate the joint posterior distribution. This algorithm operates by passing messages along edges of the factor graph to propagate evidence through the network. The factor graph model allows for the systematic update of

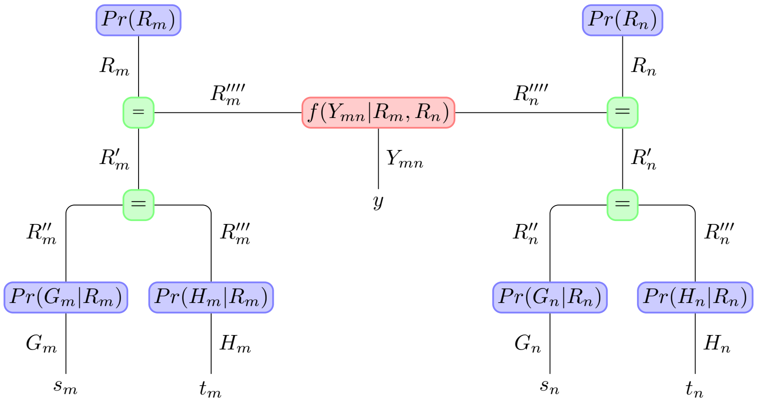

Fig. 1. Factor Graph representation for combining delay and node-local evidence to simultaneously geolocate a pair of nodes m,n

directly connected nodes, thereby reasoning about the location of a host based on the location of its directly connected hosts that may have inaccurate or missing node-local evidence.

An algorithm for automatically constructing factor graphs for arbitrary IP topologies is given in [7]. The authors applied their factor graph construction method to several test cases, applied the sum-product algorithm to find posterior distributions, and compared their results against ground truth. On both simulated and real-world test cases, they reported improved accuracy in geolocation from fusing delay data with node-local information.

IV. IP Geolocation Probabilistic Ontology

IP address geolocation is an important component capability for a broad variety of applications in cyber security and other information technology domains. Explicitly representing the the semantics of the IP address geolocation model can support model reuse across applications and interoperability with other kinds of models. For example, IP geolocation can contribute to predicting, diagnosing and responding to large-scale cyber attacks. Incorporating a geolocation capability into a cyber situation awareness and response system is facilitated by semantic awareness of the system and the component module.

As the factor graph in Fig. 1 makes clear, the geolocation model consists of modular elements that are assembled into a larger model to reason about a given network topology and patterns of available evidence. Additional evidence types can be added in a modular way by augmenting the graph with additional nodes and edges representing the new evidence. In a similar manner, the geolocation model could be used within a more comprehensive system. For example, a graphical probability model library for cyber attack plan recognition (*c.f.* [21]) could be augmented with elements representing geolocation information, which could then be referenced by attack plan models.

It is worth noting that the model of Section III does not consider intentional efforts by users to thwart attempts at geolocation. Reasons for evading geolocation are diverse, including privacy concerns, overcoming geographical restrictions on content access, and disguising the source of cyber-attacks. To address the problem of IP spoofing, a technique common in denial of service (DoS) attacks, it is important to draw a distinction between user geolocation and IP geolocation. As pointed out by [22], user geolocation seeks to identify the location of a user who requests content or attempts to connect to a specific resource, whereas IP geolocation seeks to identify the geographic location of a device given its IP address. An IP geolocation capability such as the one presented above can support user geolocation by tracing the path of a spoofed packet to the network edge and geolocating the device nearest to the origin. Other techniques exist to extract the actual IP address of the attacker, which can be geolocated using an IP geolocation method. Again, our IP geolocation could be combined with additional modular components to form a user geolocation capability.

Representing the model as a probabilistic ontology supports this kind of model interoperability and reuse. Ideally, such a probabilistic ontology would be built on an existing ontology of the cyber domain (*e.g.*, [2]). As such, many of the random variables in the model should already be represented in the ontology, and probabilistic ontology development would largely involve augmenting the existing ontology with information about uncertainties. For the purpose of illustrating the approach, we constructed a limited, partial ontology consisting of entities, properties and relationships needed to reason about IP geolocation and augmented that ontology with uncertainty information. Clearly, a comprehensive ontology of the cyber domain would represent additional general knowledge and specific domain knowledge not included here.

Our probabilisitic ontology is represented in the PR-OWL language [6] and implemented in the UnBBayes-MEBN open-source PR-OWL reasoning tool [23]. Our representation includes some workarounds to overcome limitations of the current version of UnBBayes-MEBN. These limitations will be addressed in future releases. The model encoded in the probabilistic ontology is equivalent to the IP geolocation model presented above.

Table II lists the entities in the partial ontology and properties used by the node geolocation probabilistic ontology. The ontology has four types of entity: IP nodes, regions where IP nodes can be located, probe packets for measuring link delays, and evidence items. Because this is an OWL ontology, all four types are subtypes of Thing.

TABLE II
ENTITIES AND ATTRIBUTES IN GEOLOCATION PROBABILISTIC ONTOLOGY

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

A property of an IP node is its location. A property of a region is its region ID, a unique identifier used to refer to the region. Properties of a probe packet include its starting and ending nodes. Properties of an evidence item include the IP node to which it refers for node local evidence, the content of a GeoIP query response, the content of a hostname lookup result, the probe packet measured by a link delay report, and the measured delay for a probe packet sent across a link.

Table III shows the relationships represented in the ontology. The entity types participating in the relationship are shown. The IsA relationship relates an entity and a type if the entity is of the given type. The ontology includes the relationships NodeDistance and RegionDistance to represent the distance between nodes and regions, respectively. Ideally, there would be only one Distance attribute to represent the distance between two spatial entities. However, the current UnBBayes-MEBN implementation does not yet support polymorphism; this capability is slated for the next release. Thus, the ontology uses two different terms to accommodate the limitations of the reasoning tool.

Fig. 2 shows the Node Geolocation probabilistic ontology. The probabilistic ontology consists of five MFrags (Multi-Entity Bayesian Network Fragments). Each MFrag defines a local probability distribution for its resident random variables, depicted by yellow ovals, conditional on their parents in the MFrag. The context random variables, depicted by green pentagons, represent conditions that must be satisfied for the local distribution definitions to be meaningful. Finally, the gray trapezoids are input random variables, which are parents of resident random variables whose distribution is defined in another MFrag.

TABLE III
RELATIONSHIPS IN GEOLOCATION PROBABILISTIC ONTOLOGY

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)

The random variables in the MFrags define a joint probability distribution over properties and relationships in the ontology. A random variable with a single argument corresponds to a property, and a random variable with two arguments corresponds to a relationship. The arguments are placeholders (called ordinary variables to distinguish them from random variables) that can be filled in by the identifiers of individuals of the appropriate types. For example, if N1 and N2 are individuals of the IPNode type, the random variable $\mathtt{Location(N1)}$ represents the uncertain location of N1, and $\mathtt{NodeDistance(N1,N2)}$ corresponds to the distance between IP nodes N1 and N2. The second argument of the IsA random variable is always a type name, indicating the type of its first argument. Thus, $\mathtt{IsA(N1,IPNode)}$ has value *True* and $\mathtt{IsA(N1,Region)}$ has value *False*. Multiple instances of these MFrags can be constructed by filling in the ordinary variables with different entity instances. The MFrag instances can then be assembled into a Bayesian network called a situation-specific Bayesian network, or SSBN.

The MFrags and local distributions are described as follows.

- *Node Location*: This MFrag defines a distribution for the *Location* random variable, representing the region in which an IP node is located. This random variable corresponds to the Location property of the IPNode entity. It also corresponds to the random variable R_n in the factor graph of Fig. 1. Its possible values are regions and it is given a uniform distribution, meaning that all regions are equally likely locations for any given node.
- *Distance*: This MFrag defines distances between regions. The *RegionDistance* random variable is initialized to a uniform distribution (or a Gaussian distribution with mean zero and very large variance). When region instances are defined, their respective *RegionDistance* random variables are set to the actual distance between each pair of regions. The *NodeDistance* random variable has a deterministic distribution, being equal to the distance between the regions in which its arguments are located. These random variables define distributions for the *RegionDistance* and *NodeDistance* relationships from Table III.
- *Probe Packet Definition*: This MFrag defines random variables *StartingNode* and *EndingNode* for probe packets sent across links. The distributions are initialized as uniform. When a delay measurement is received, they are set to the starting and ending node for the the probe packet. These random variables define distributions for the StartingNode and EndingNode properties of a

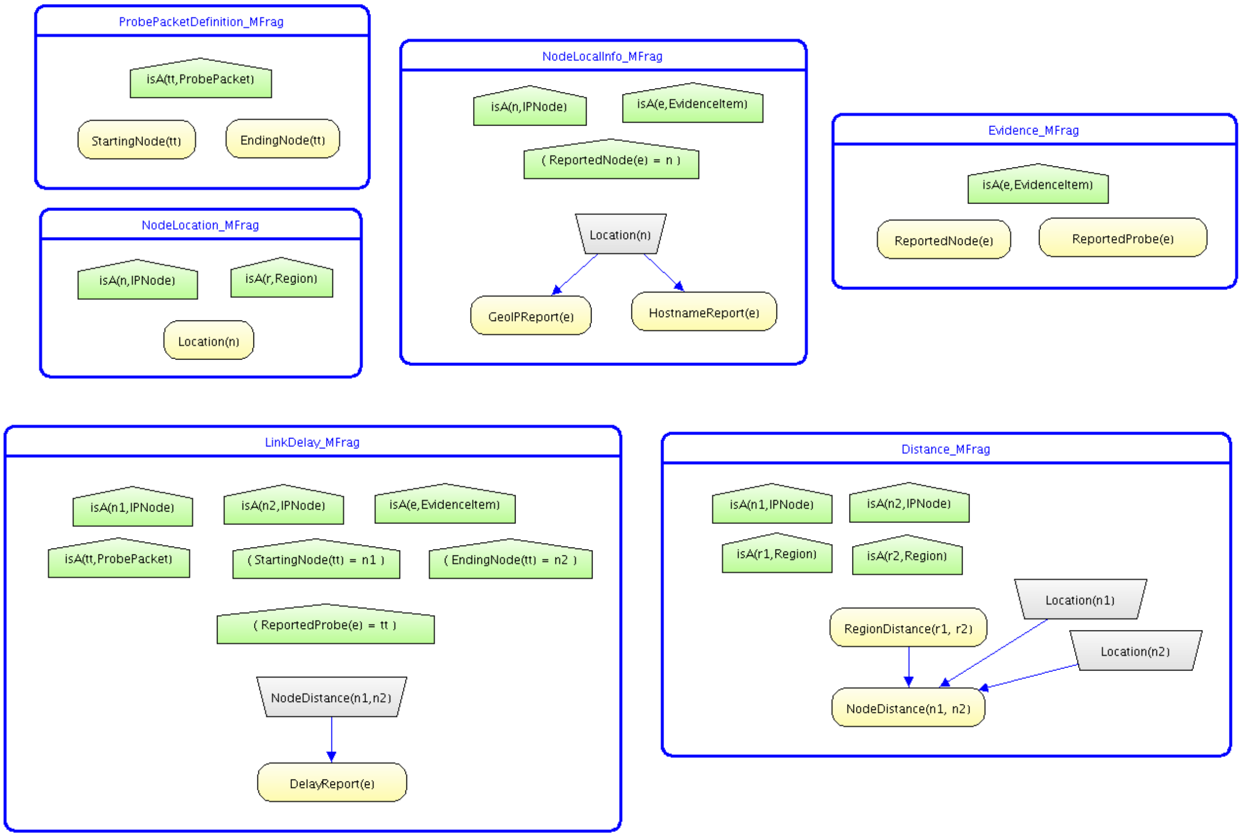

Fig. 2. Node Geolocation Probabilisitic Ontology

ProbePacket entity as defined in Table II.

- *Evidence*: This MFrag defines random variables *ReportedNode* and *ReportedProbe* for evidence items. The former defines the IP node to which a GeoIP query or a hostname lookup refers. The latter defines the probe packet to which a given link delay measurement refers. These random variables are initialized to uniform distributions and are set to the appropriate values when reports are received. These random variables define distributions for the ReportedNode and ReportedProbe properties of EvidenceItem entities as defined in Table II.
- *Link Delay*: This MFrag defines the distribution for a link delay measurement conditional on the distance between the starting node and ending node for the corresponding probe packet. The *DelayReport* random variable corresponds to the random variable Y_{mn} in the factor graph of Fig. 1. It has the normal distribution given by (4), or a mixture of normal distributions if router delay is being considered in the model. This random variable defines the distribution for the DelayReport property of an EvidenceItem entity as presented in Table II.
- *Node Local Information*: This MFrag represents evidence local to a given node. The resident nodes *GeoIPReport* and *HostnameReport* correspond to the random variables G_n and H_n, respectively, in the factor graph of Fig. 1. The local distributions for *GeoIPReport* and *HostnameReport* are given by (2) and (3), respectively. These random variables define distributions for the GeoIPReport and HostnameReport properties of an IPNode entity shown in Table II.

The probabilistic ontology is applied to a given network topology and set of measurements as follows. Assume that we are given a set of nodes, a set of regions, a network topology defining node connectivity, link delay measurements for nodes connected by the topology, and GeoIP query and hostname lookup results for some or all of the nodes. Inference about node locations proceeds as follows.

1) Create an instance of Region for each region. Define the regions as mutually exclusive. Give each region an ID, and set the value of RegionID to the region's ID. For each pair of regions, set the value of RegionDistance to the distance between the regions.
2) Create an instance of IPNode for each node in the network. Define the nodes as mutually exclusive.
3) Create an instance of ProbePacket for each probe packet for which the propagation delay has been measured. Set the properties StartingNode and EndingNode to the instances of IPnode corresponding

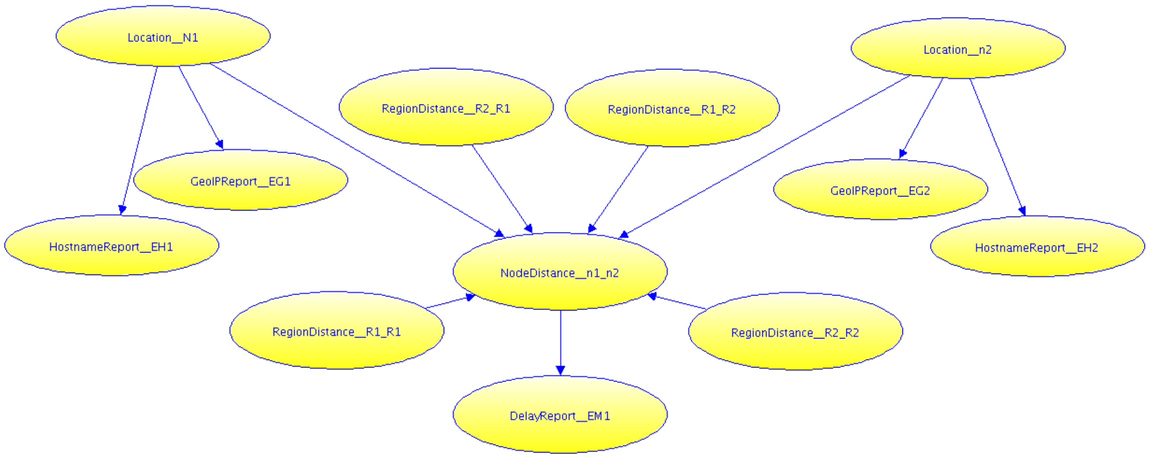

Fig. 3. Constructed Bayesian network for geolocating a pair of IP nodes

to the endpoints of the link.

4) Create an instance of `EvidenceItem` for each GeoIP query result. For each report, set the property `ReportedNode` to the `IPnode` instance to which the report refers, and the property `GeoIPReport` to the region indicated by the report.
5) Create an instance of `EvidenceItem` for each hostname lookup result. For each report, set the property `ReportedNode` to the `IPnode` instance to which the report refers, and the property `HostnameReport` to the region indicated by the report.
6) Create an instance of `EvidenceItem` for each link delay measurement. For each report, set the property `ReportedProbe` to the instance of `ProbePacket` to which the report refers. Set the property `DelayReport` to the measured delay across the link.
7) Run a query to find the posterior distribution of the `NodeLocation` properties. This involves assembling a situation-specific Bayesian network containing the random variable instances created in the above steps. The MFrags containing the random variable instances are retrieved and instantiated, and then combined by unifying on common random variables. The result is a Bayesian network to reason about IP node locations.

Fig. 3 shows the situation-specific Bayesian network produced by the UnBBayes software for the case of a network with two nodes connected by a single link. The Bayesian network of Fig. 3 and the factor graph of Fig. 1 encode the identical joint distribution for node locations and evidence. The model of (5) and the probabilistic ontology of Fig. 2 encode formally equivalent joint distributions over node locations, node local evidence and transmission delay evidence.

The SSBN of Fig. 3 is a hybrid Bayesian network containing both discrete and continuous random variables. The distance random variables are real-valued and continuous; the other random variables are discrete. The graph of Fig. **??** is a polytree, and exact inference is possible using an algorithm for conditional linear Gaussian (CLG) Bayesian networks. For larger networks with complex network topologies, the graph will contain cycles and exact inference is intractable. Approximate inference algorithms such as the one presented in [24] can be applied.

V. Conclusion

Explicitly representing domain semantics in computable form supports maintainability, interoperability and extensibility of systems. For problems characterized by reasoning under uncertainty, a semantically rich representation for sources of uncertainty should be appropriately integrated with the domain ontology. This paper presented a case study of a probabilistic ontology for large-scale IP address geolocation. The probabilistic ontology integrates an existing factor graph model for IP geolocation with a domain ontology representing geolocation knowledge. Random variables in the factor graph model correspond to uncertain properties and relationships in the domain ontology. The model is represented as a PR-OWL probabilistic ontology that augments an OWL domain ontology by defining probability distributions for uncertain properties and relationships. Reasoning with the probabilistic ontology is performed by creating instances of the relevant entities, instantiating copies of the random variables by filling in their arguments with appropriate entity instances, and assembling them into a Bayesian network to reason about the particular problem instance. The model can be used to reason about arbitrary numbers of IP nodes and regions, arbitrary network topologies, and arbitrary numbers of evidence items.

Acknowledgment

The authors thank Shou Matsumoto for assisting with development of the UnBBayes representation of the probabilistic ontology.

References

[1] A. Kott, "Towards Fundamental Science of Cyber Security," in *Network Science and Cybersecurity*, R. E. Pino, Ed., New York, 2014, vol. 55.

[2] A. Oltramari, L. Cranor, R. Walls, and P. McDaniel, "Building an Ontology of Cyber Security," in *Proceedings of the Ninth Conference on Semantic Technologies for Intelligence, Defense, and Security (STIDS 2014)*, ser. CEUR Workshop Proceedings, K. B. Laskey, I. Emmons, and P. C. G. Costa, Eds. Aachen: George Mason University, 2014, pp. 54–61. [Online]. Available: http://ceur-ws.org/Vol-xxx/

[3] R. Dipert, "The Essential Features of an Ontology for Cyberwarfare," in *Conflict and Cooperation in Cyberspace - The Challenge to National Security,*, P. A. Yannakogeorgos and A. B. Lowther, Eds. Taylor and Francis, 2013.

[4] B. Barnett and A. Crapo, "A Semantic Model for Cyber Security," in *Proceedings of the Fifth Grid-Interop Forum*. Gridwise Architectural Council, 2011. [Online]. Available: http://www.gridwiseac.org/pdfs/forum_papers11/barnett_paper_gi11.pdf

[5] N. F. Noy and D. L. Mcguinness, "Ontology development 101: A guide to creating your first ontology," Knowledge Systems Laboratory, Stanford University, Tech. Rep., 2001.

[6] K. Laskey, P. Costa, and T. Janssen, "Probabilistic ontologies for knowledge fusion," in *2008 11th International Conference on Information Fusion*, Jun. 2008, pp. 1–8.

[7] S. Chandekar and B.-P. Paris, "Large-scale, discrete IP geolocation via multi-factor evidence fusion using factor graphs," in *18th International Conference on Information Fusion*, Jul. 2015.

[8] R. Koch, M. Golling, and G. D. Rodosek, "Advanced Geolocation of IP Addresses," in *International Conference on Communication and Network Security (ICCNS)*, 2013, pp. 1–10. [Online]. Available: http://www.waset.org/publications/16111

[9] M. Crovella and B. Krishnamurthy, *Internet Measurement infrastructure, traffic and applications*. England: John Wiley and Sons, 2006.

[10] MaxMind LLC, "GeoIP." [Online]. Available: http://www.maxmind.com

[11] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for internet hosts," in *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 4. ACM, 2001, pp. 173–185.

[12] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1219–1232, 2006.

[13] A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. Duarte, "Improving the accuracy of measurement-based geographic location of internet hosts," *Computer Networks*, vol. 47, no. 4, pp. 503–523, 2005.

[14] M.J.Arif, S.Karunasekara, S.Kulkarni, A.Gunatilaka, and B.Ristic, "Internet host geolocation using maximum likelihood estimation technique," *IEEE International Conference on Advanced information Networking and Applications*, 2010.

[15] I. Youn, B. L. Mark, and D. Richards, "Statistical geolocation of internet hosts," in *Computer Communications and Networks, 2009. ICCCN 2009. Proceedings of 18th Internatonal Conference on*. IEEE, 2009, pp. 1–6.

[16] DIMES, "Ip topology," http://www.netdimes.org.

[17] Y. Shavitt and E. Shir, "Dimes: Let the internet measure itself," *ACM SIGCOMM Computer Communication Review*, vol. Vol 35, October 2005.

[18] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[19] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, 2007.

[20] G. D. Forney Jr, "Codes on graphs: normal realizations," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 520–548, 2001.

[21] C. W. Geib and R. P. Goldman, "A probabilistic plan recognition algorithm based on plan tree grammars," *Artificial Intelligence*, vol. 173, no. 11, pp. 1101–1132, Jul. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370209000459

[22] J. A. Muir and P. C. V. Oorschot, "Internet geolocation: Evasion and counterevasion," *ACM Computing Surveys (CSUR)*, vol. 42, no. 1, p. 4, 2009.

[23] P. C. G. d. Costa, M. Ladeira, R. N. Carvalho, K. B. Laskey, L. L. Santos, and S. Matsumoto, "A First-Order Bayesian Tool for Probabilistic Ontologies," in *FLAIRS Conference*, 2008, pp. 631–636.

[24] W. Sun, K.-C. Chang, and K. Laskey, "Scalable inference for hybrid Bayesian networks with full density estimations," in *2010 13th Conference on Information Fusion (FUSION)*, Jul. 2010, pp. 1–8.