Towards Sound Analysis of Computer Evidence

Imani Palmer

Professor Roy Campbell
University of Illinois at Urbana-Champaign
Department of Computer Science

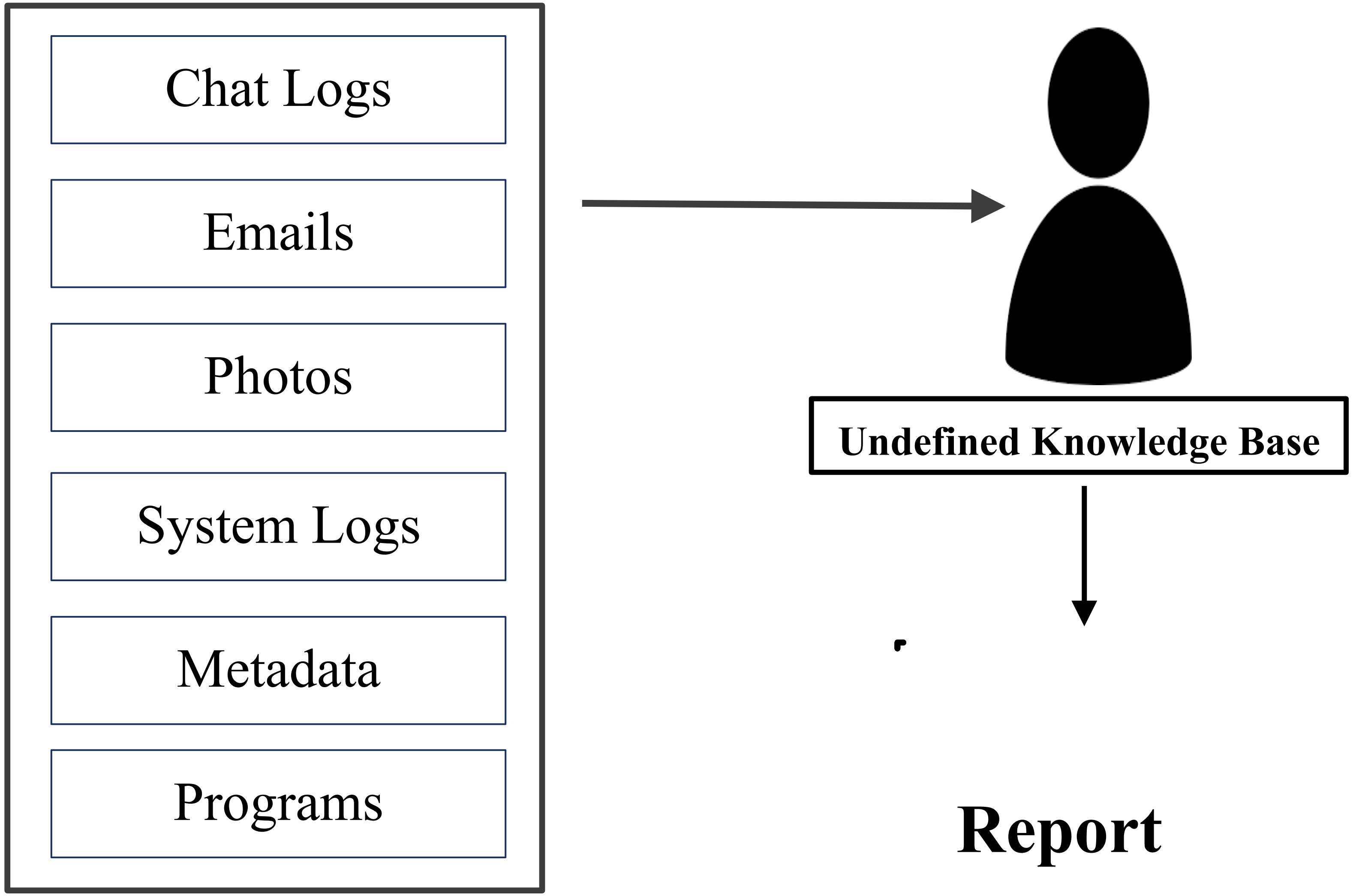
Boris Gelfand, PhD
Los Alamos National Laboratory
Advanced Research in Cyber Systems

Acquisition

Analysis

Reporting

GUILTY



Chat Logs

Emails

Photos

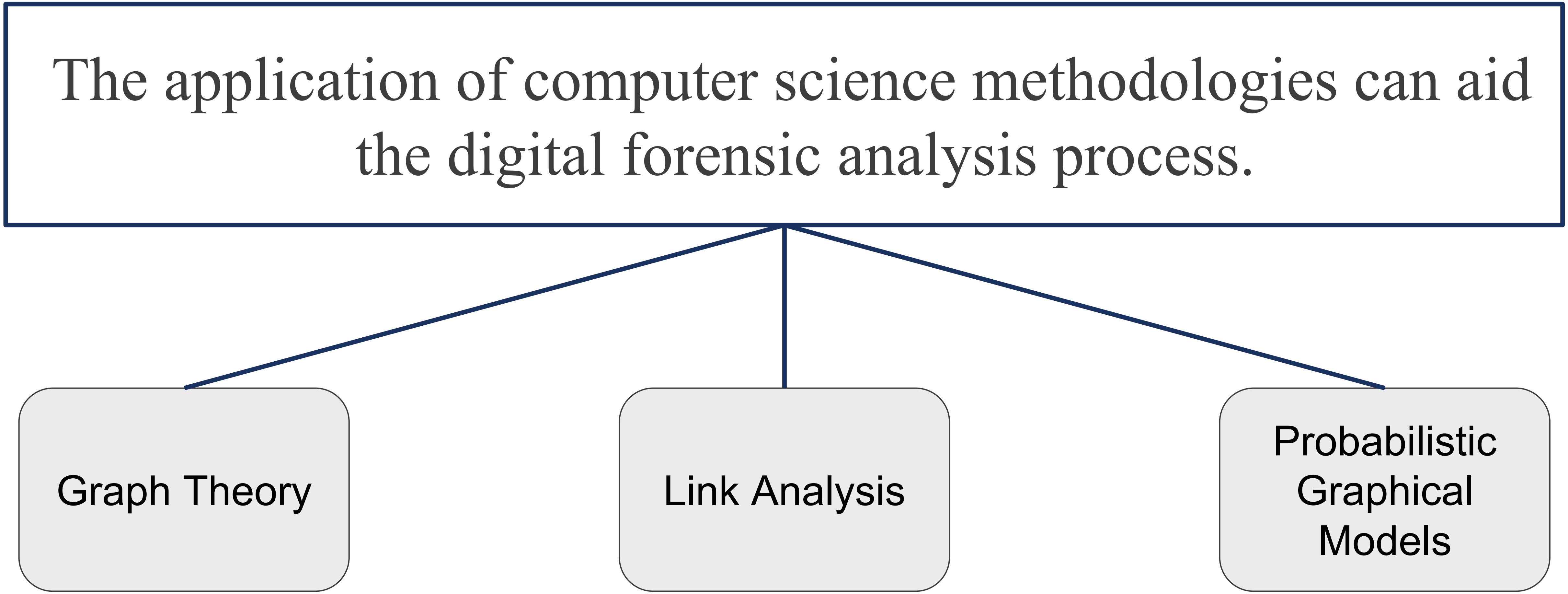
System Logs

Metadata

Programs

Undefined Knowledge Base

Report



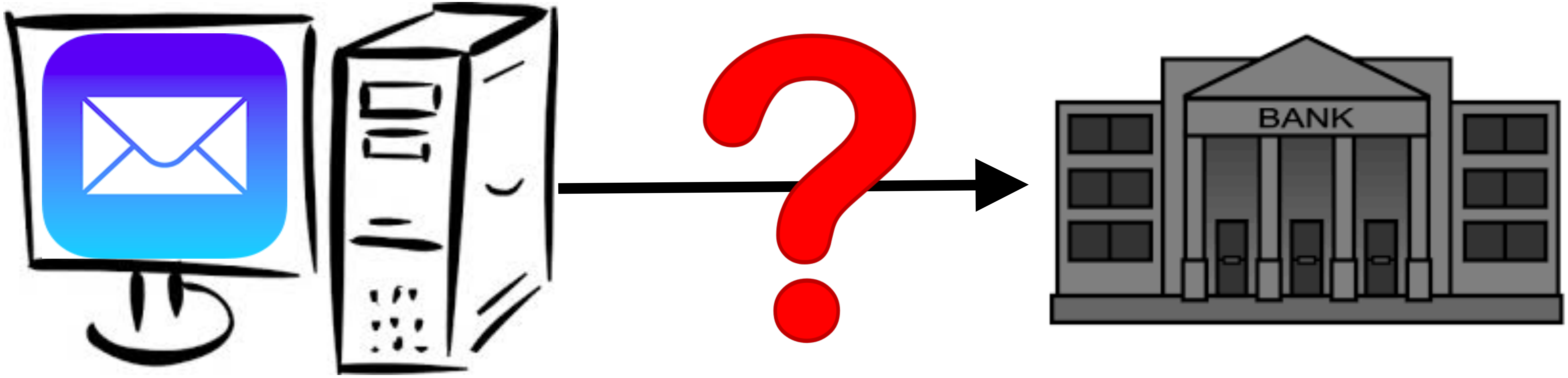
The application of computer science methodologies can aid the digital forensic analysis process.

Graph Theory

Link Analysis

Probabilistic Graphical Models

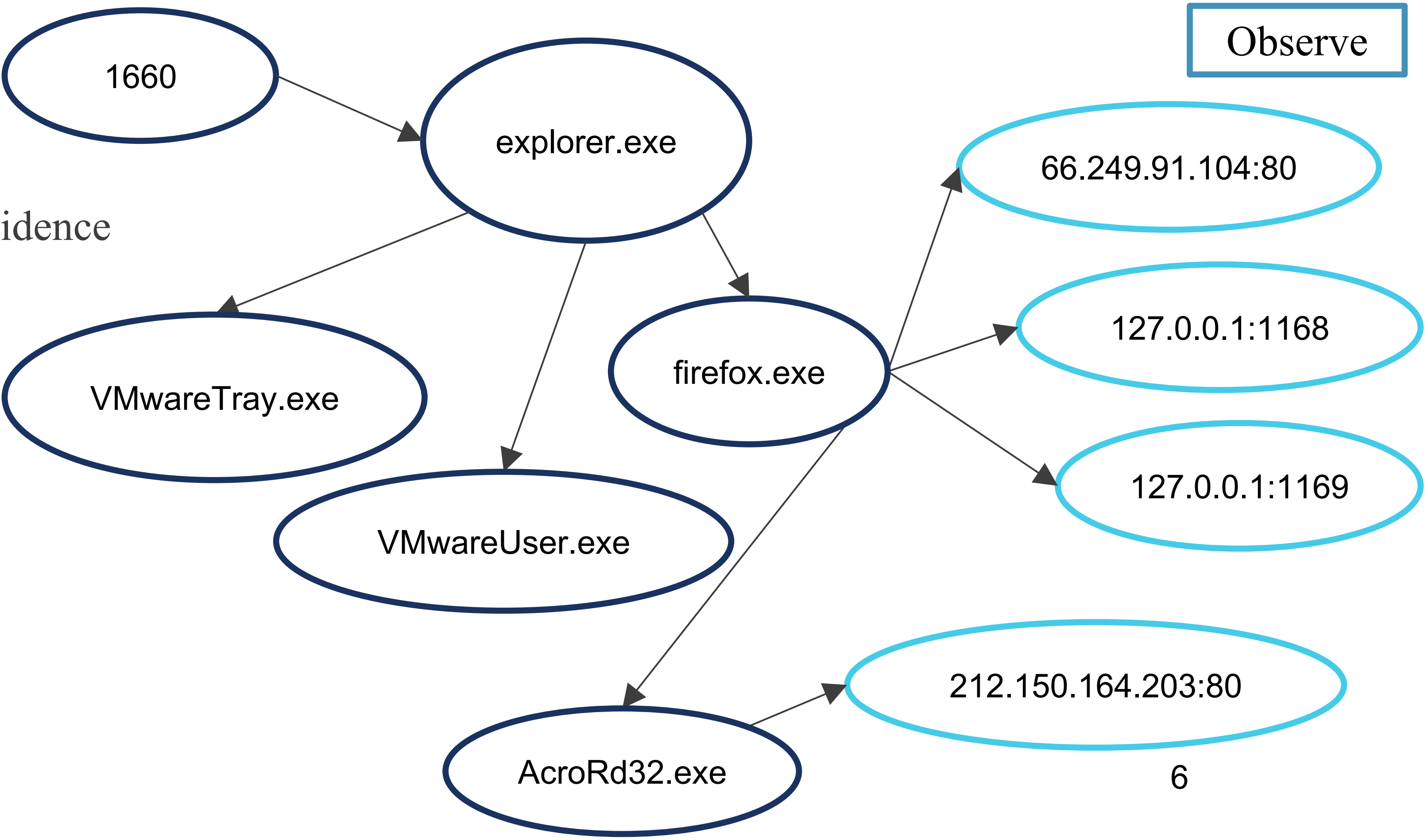
Case Study



Observe Evidence

- Graph-based representation of the evidence
  - Nodes
    - Processes and sockets
  - Edges
    - Parent process -> Child Process
    - Process -> Socket

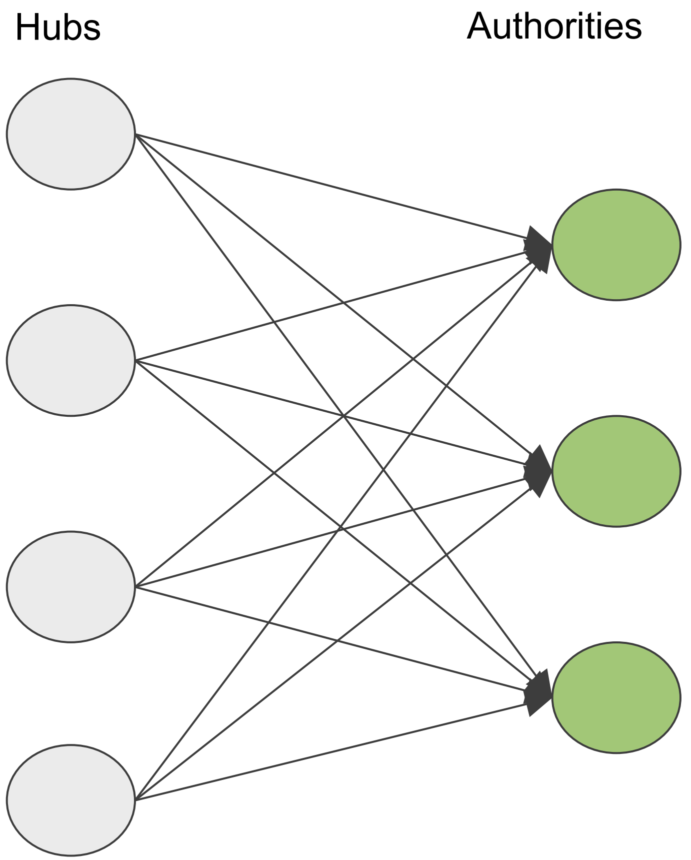
Observe



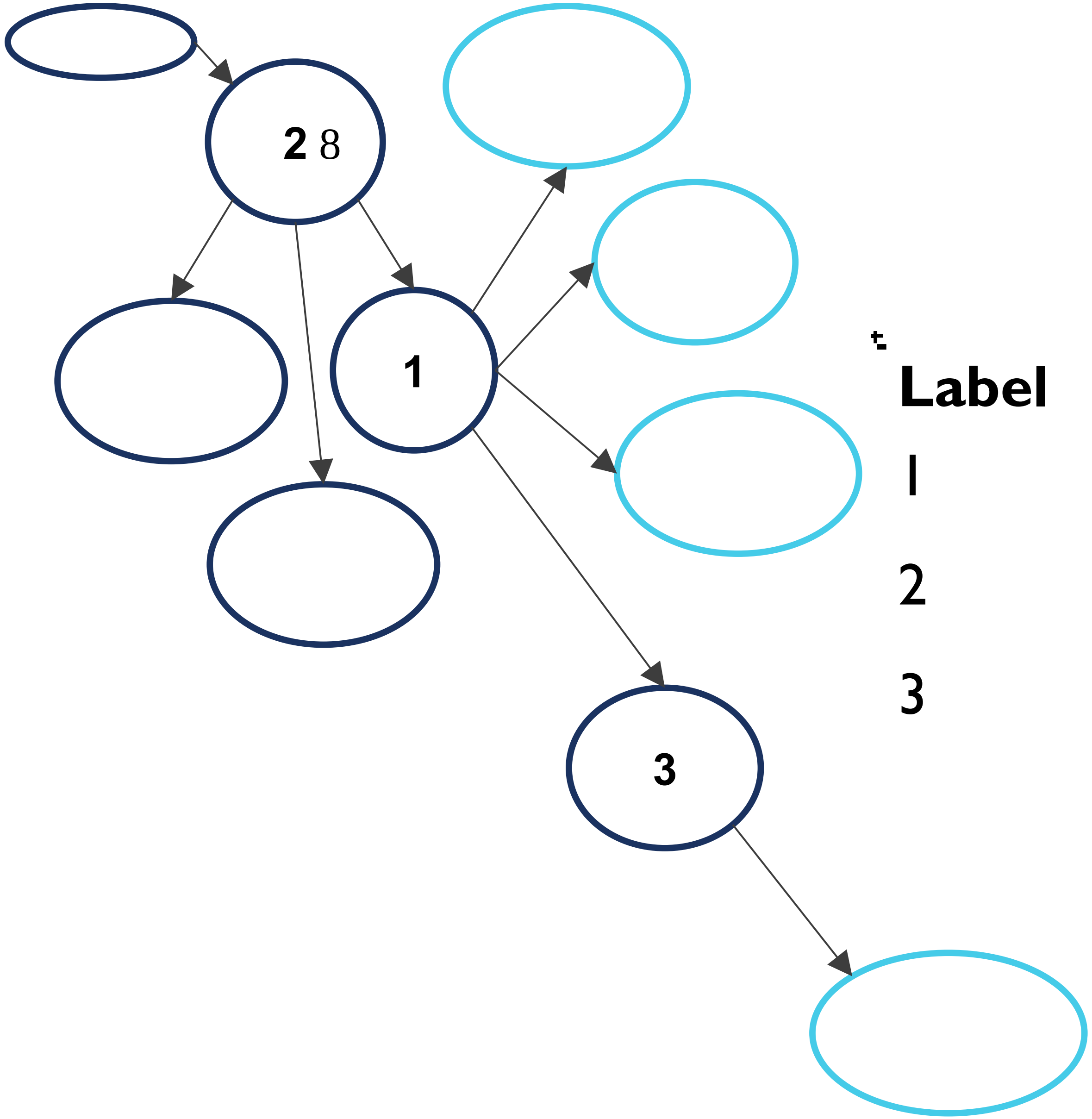
Hyperlink-Induced Topic Search (HITS)

Observe

- **Authority:** a node that hubs link to
- **Hub::** a node that links to many authorities



Observe

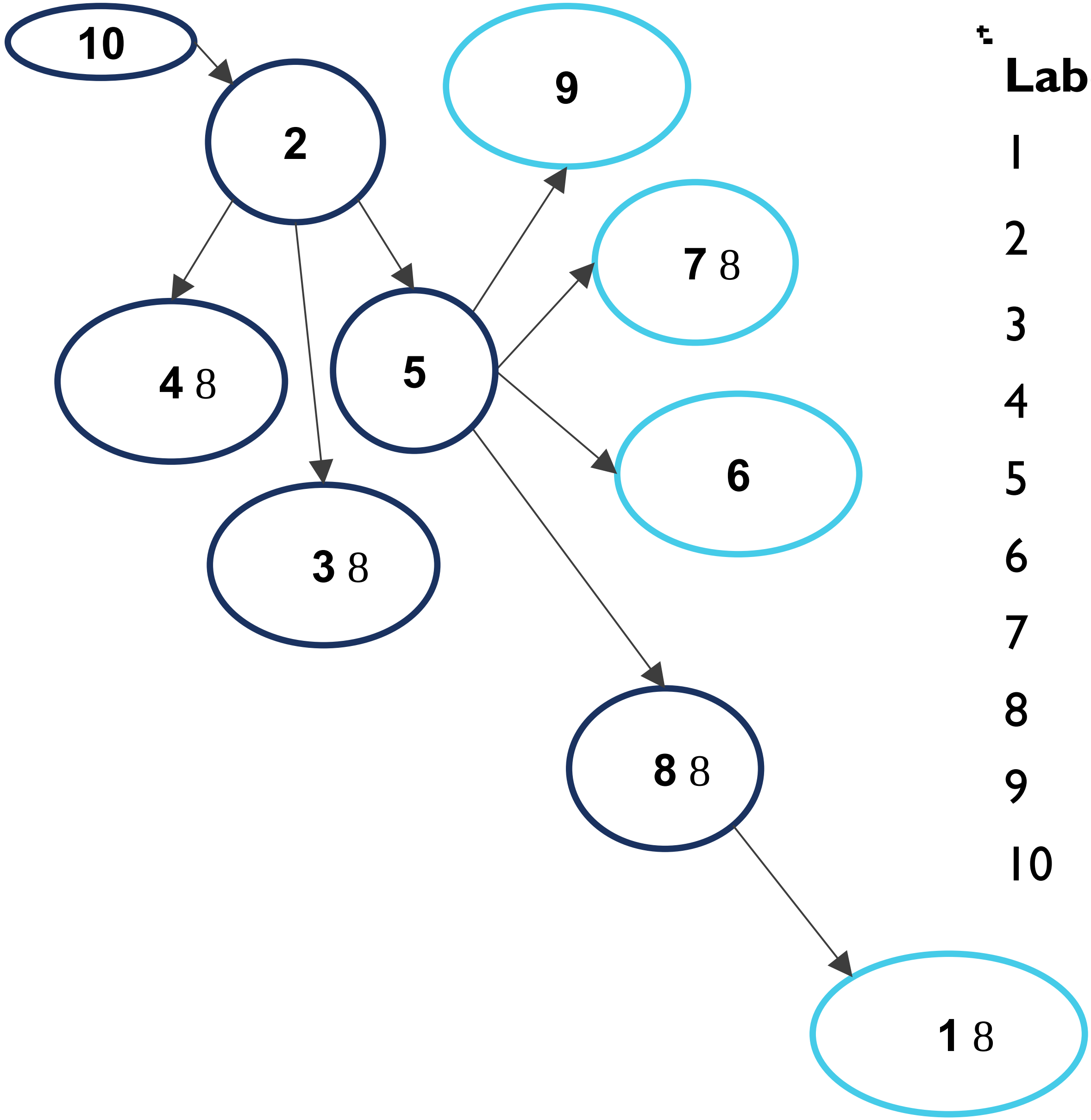


Label	Node	Hub	Authority
1	firefox.exe	0.9999	7.9728e-09
2	explorer.exe	2.3918e-08	1.8807-e37
3	AcroRd32.exe	1.8807e-37	0.2499

PageRank

Observe

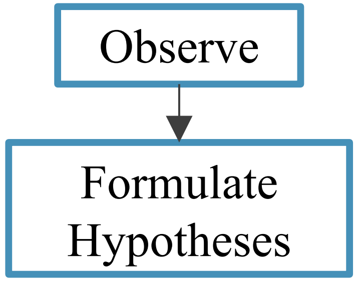
- Relationship from **Node A** to **Node B** is a vote for **Node B** cast by **Node A**
- Votes cast by nodes that are important weigh more heavily
- Numeric value that represents the importance of a node present on a graph



Observe

Label	Node	PageRank
1	212.150.164.203:80	0.1431
2	explorer.exe	0.1246
3	VMwareUser.exe	0.1026
4	VMwareTray.exe	0.1026
5	firefox.exe	0.1026
6	127.0.0.1:1169	0.0891
7	127.0.0.1:1168	0.0891
8	AcroRd32.exe	0.0891
9	66.249.91.104:80	0.0891 *
10	1660	0.0673 *

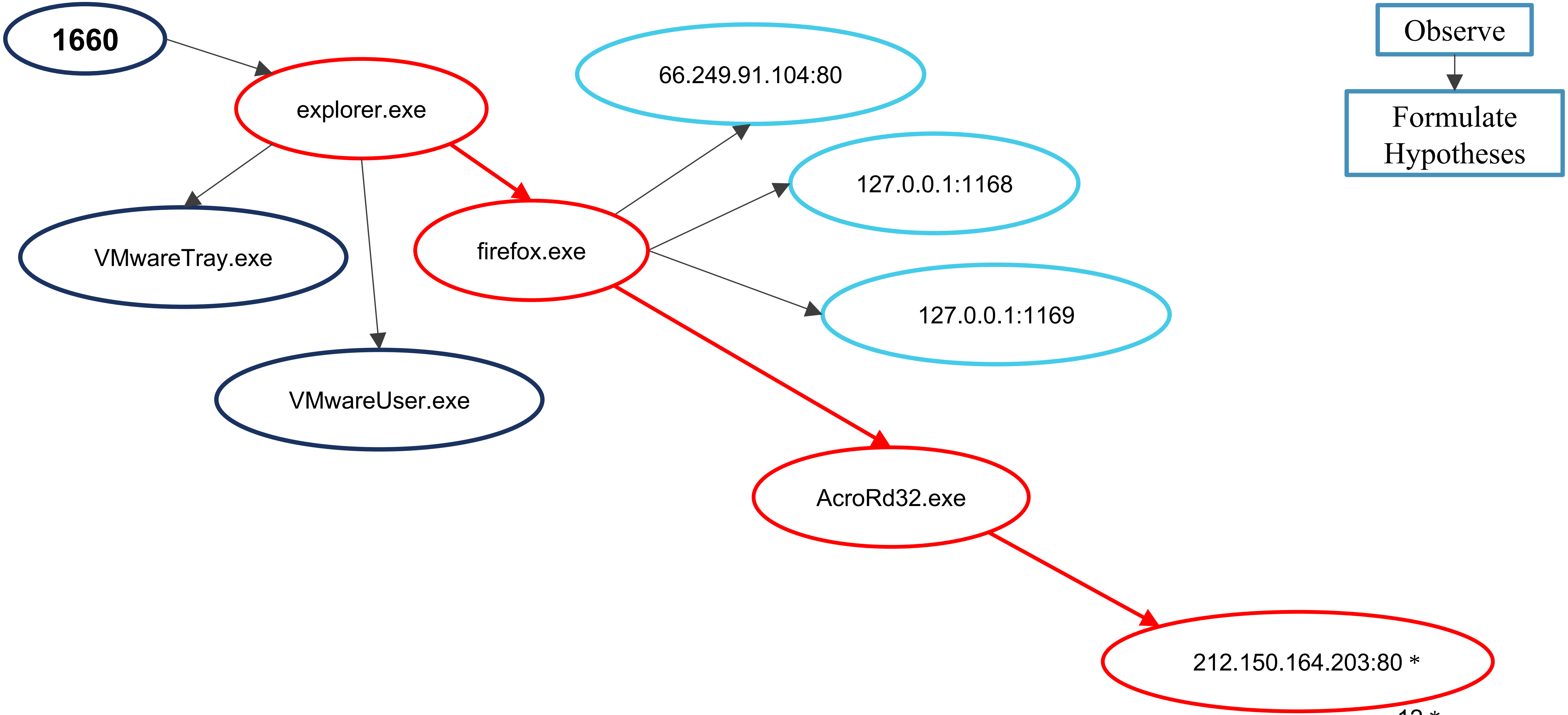
Formulate Hypotheses



Observe

Formulate Hypotheses

- Determine that hypothesis ***H*** is supported by a chain of evidence
- Graph traversal
- **Hypothesis**: X downloaded a file that made a network connection



Observe

Formulate Hypotheses

1660

explorer.exe

VMwareTray.exe

VMwareUser.exe

firefox.exe

66.249.91.104:80

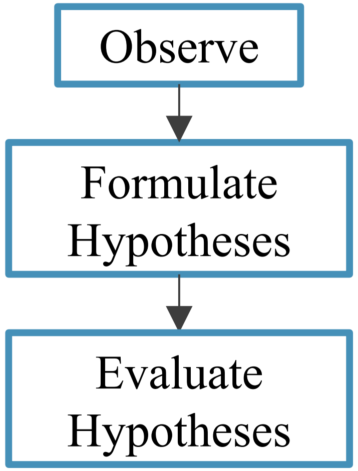
127.0.0.1:1168

127.0.0.1:1169

AcroRd32.exe

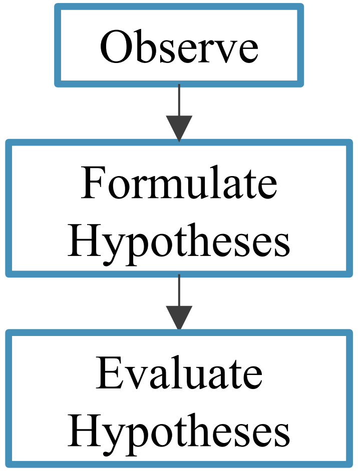
212.150.164.203:80 *

Evaluate Hypotheses



- Test abductive reasoning
- Reason about hypotheses
- Uncertainty
- Probabilistic approaches

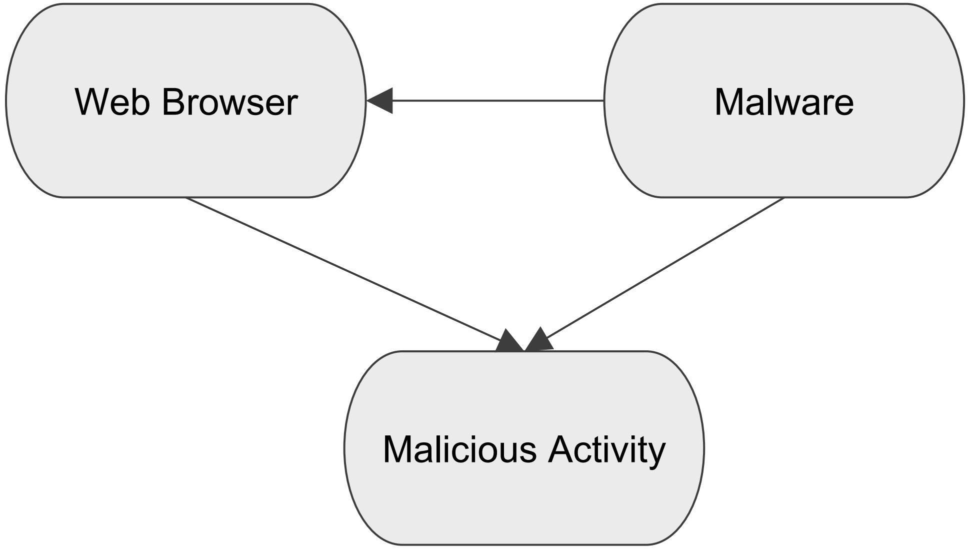
Bayesian Network


Observe

Formulate Hypotheses

Evaluate Hypotheses

- Probabilistic graphical model
- Represent evidence & conditional dependencies via a DAG
- Compute probabilities

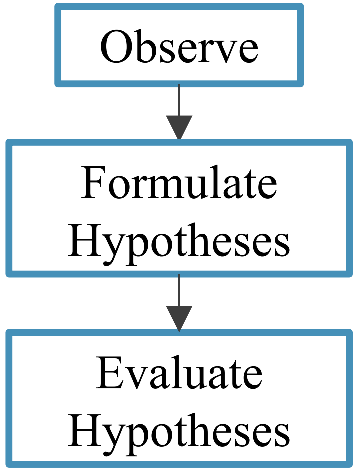

Web Browser

Malware

Malicious Activity

Bayes’ Theorem

- A method to calculate the probability of a hypothesis


Observe

Formulate Hypotheses

Evaluate Hypotheses

$$P(H|E) = P(H)\ P(E|H)\ /\ P(E)$$

P(H): prior probability of hypothesis *H*

P(E): prior probability of evidence *E*

P(*H*|*E*): probability of *H* given *E*

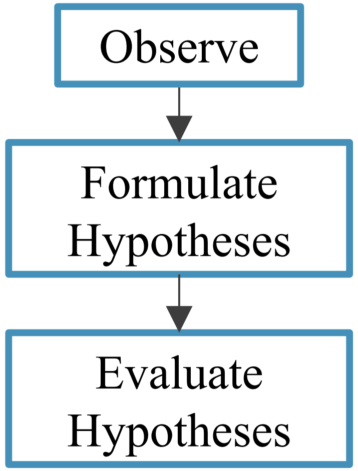
P(*E*|*H*): probability of *E* given *H*

$$\textbf{likelihood ratio} = \frac{\textbf{posterior probability x normalizing constant}}{\textbf{hypothesis prior probability}}$$

Building the Model

- H: User X downloaded a malicious file onto their computer

Node	State	P(H)
H	Yes	0.333
	No	0.333
	Uncertain	0.333

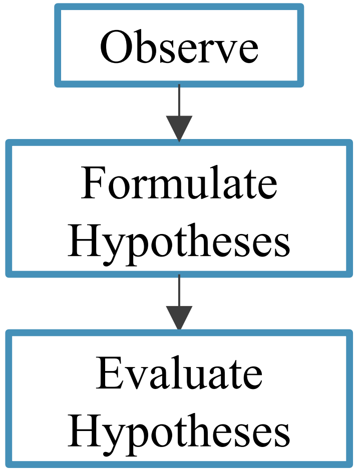


Observe

Formulate Hypotheses

Evaluate Hypotheses

Determine Informative Prior Probabilities



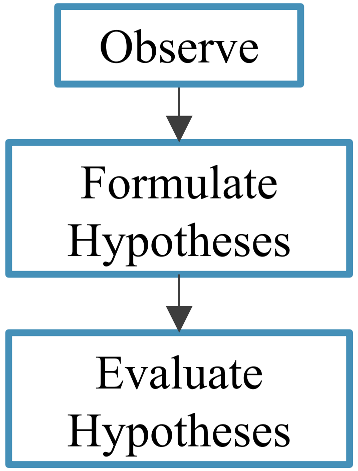
Observe

Formulate Hypotheses

Evaluate Hypotheses

- Previous approaches have relied on uninformative priors
  - An investigator can determine priors
- Informative prior
  - Survey investigators to inform the priors
  - Probability mass function

Informative Priors

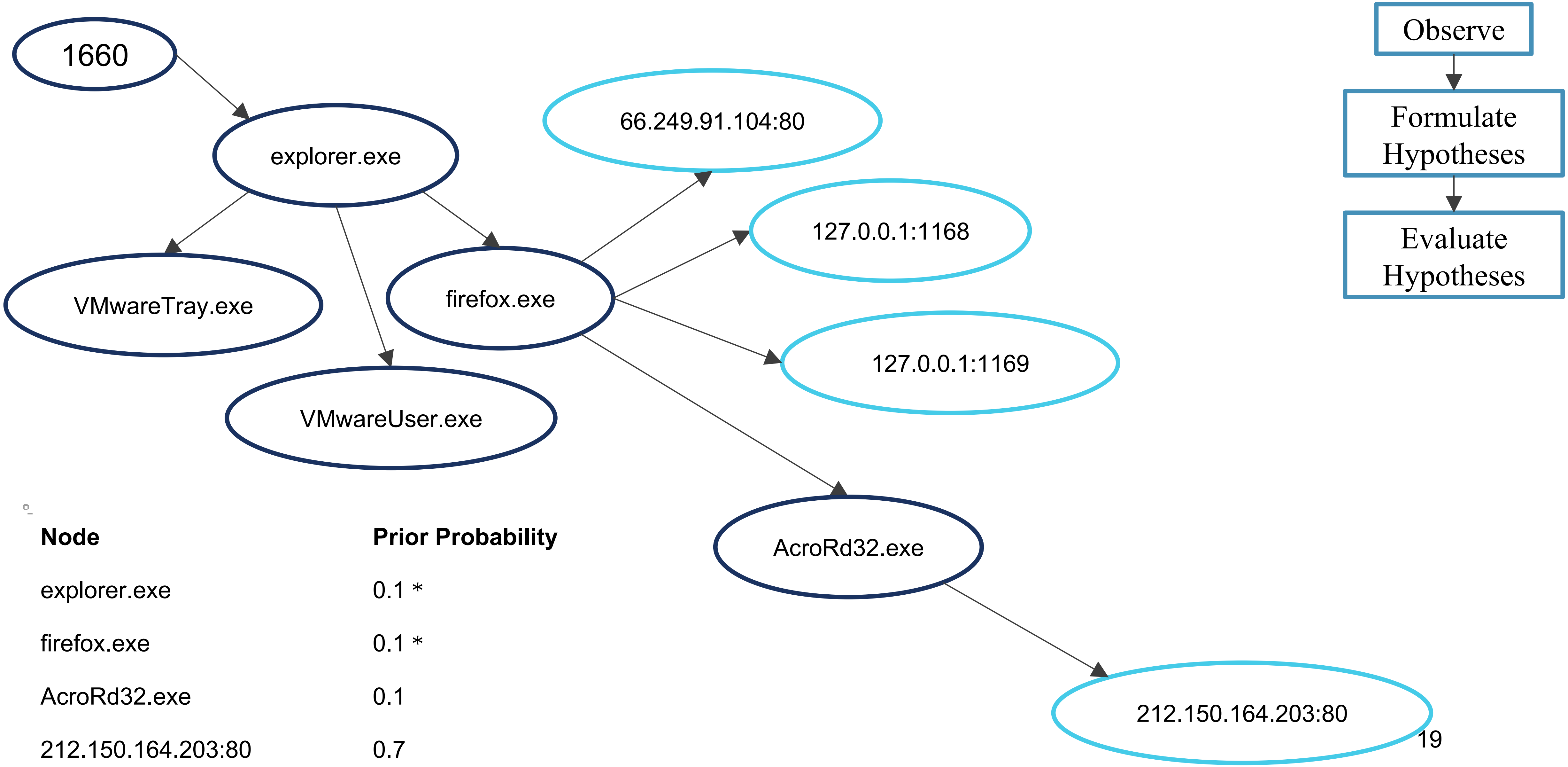

Observe

Formulate Hypotheses

Evaluate Hypotheses

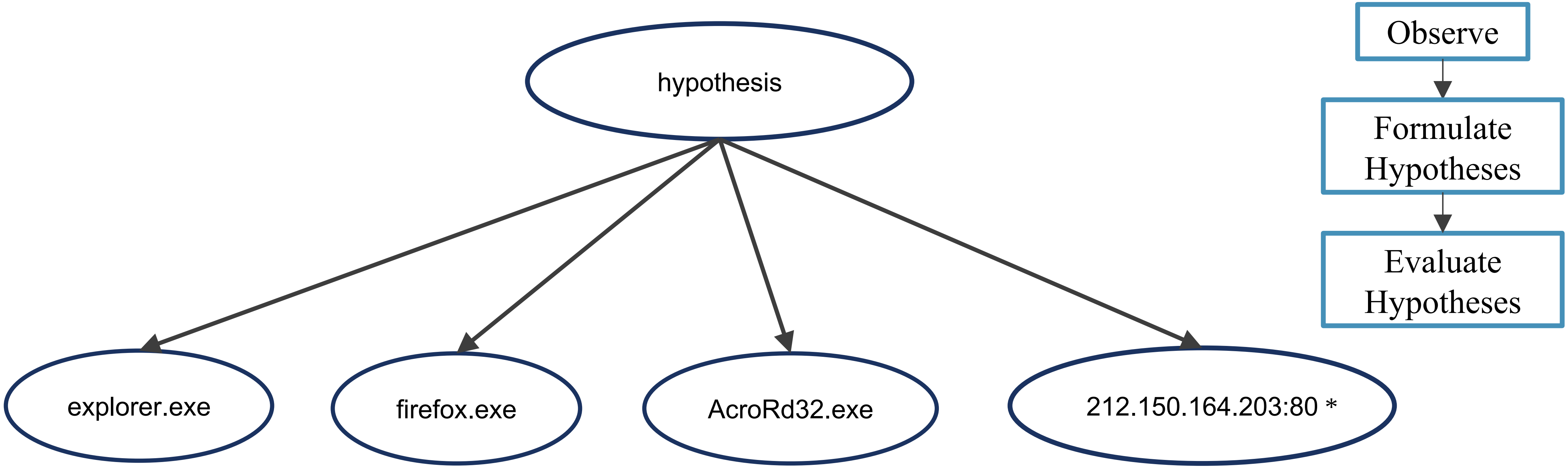
- Degree distribution
- Probability mass function

Degree	Degree Distribution	Probability
1	7	0.7
2	1	0.1
4	1	0.1
5	1	0.1

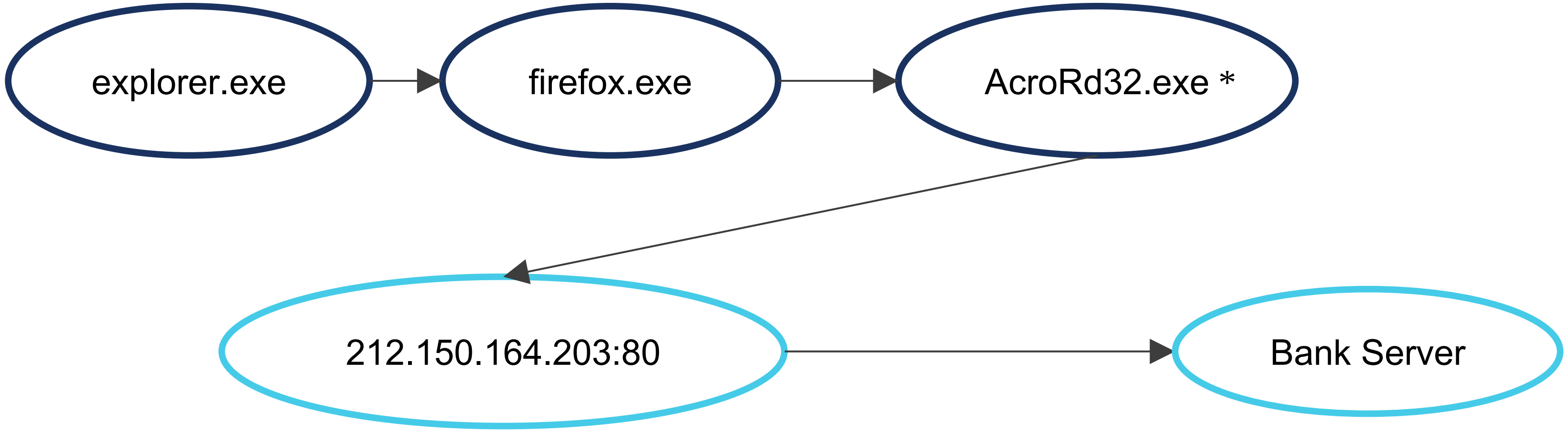


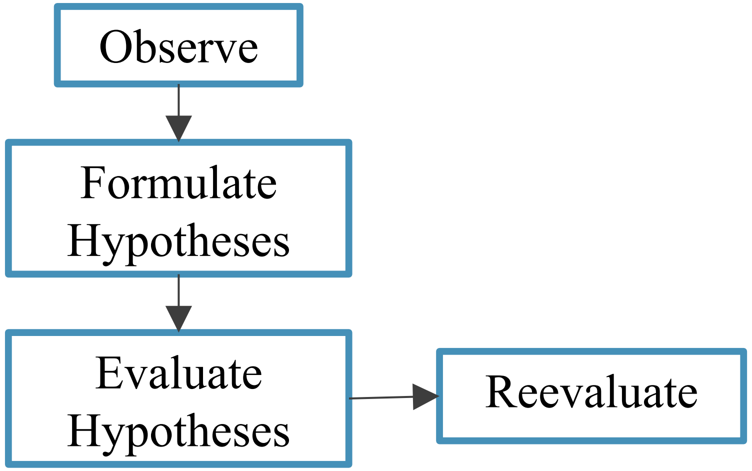
Node	**Prior Probability**
explorer.exe	0.1 *
firefox.exe	0.1 *
AcroRd32.exe	0.1
212.150.164.203:80	0.7

Evaluate Hypotheses



Yes	No	Uncertain
0.33050901	0.48902865	0.84046234

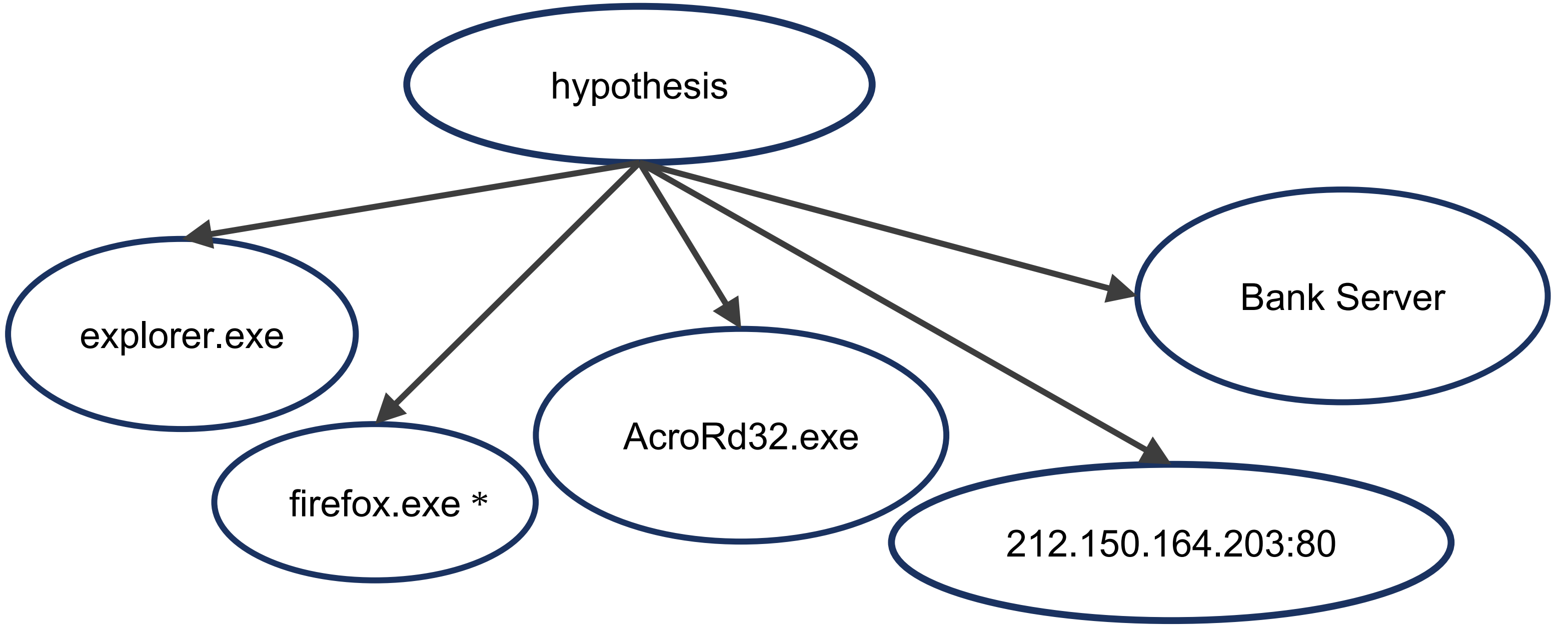


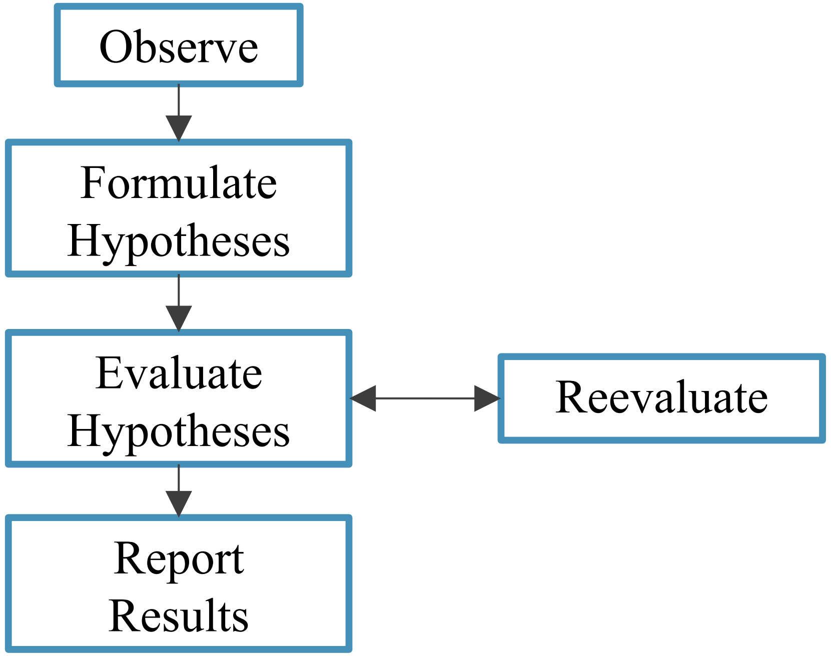


Degree	Degree Distribution	Probability
1	2	0.4
2	3	0.6

Node	Probability
explorer.exe	0.4
firefox.exe	0.6
AcroRd32.exe	0.6
212.150.164.203:80	0.6
Bank Server	0.6

Report Results





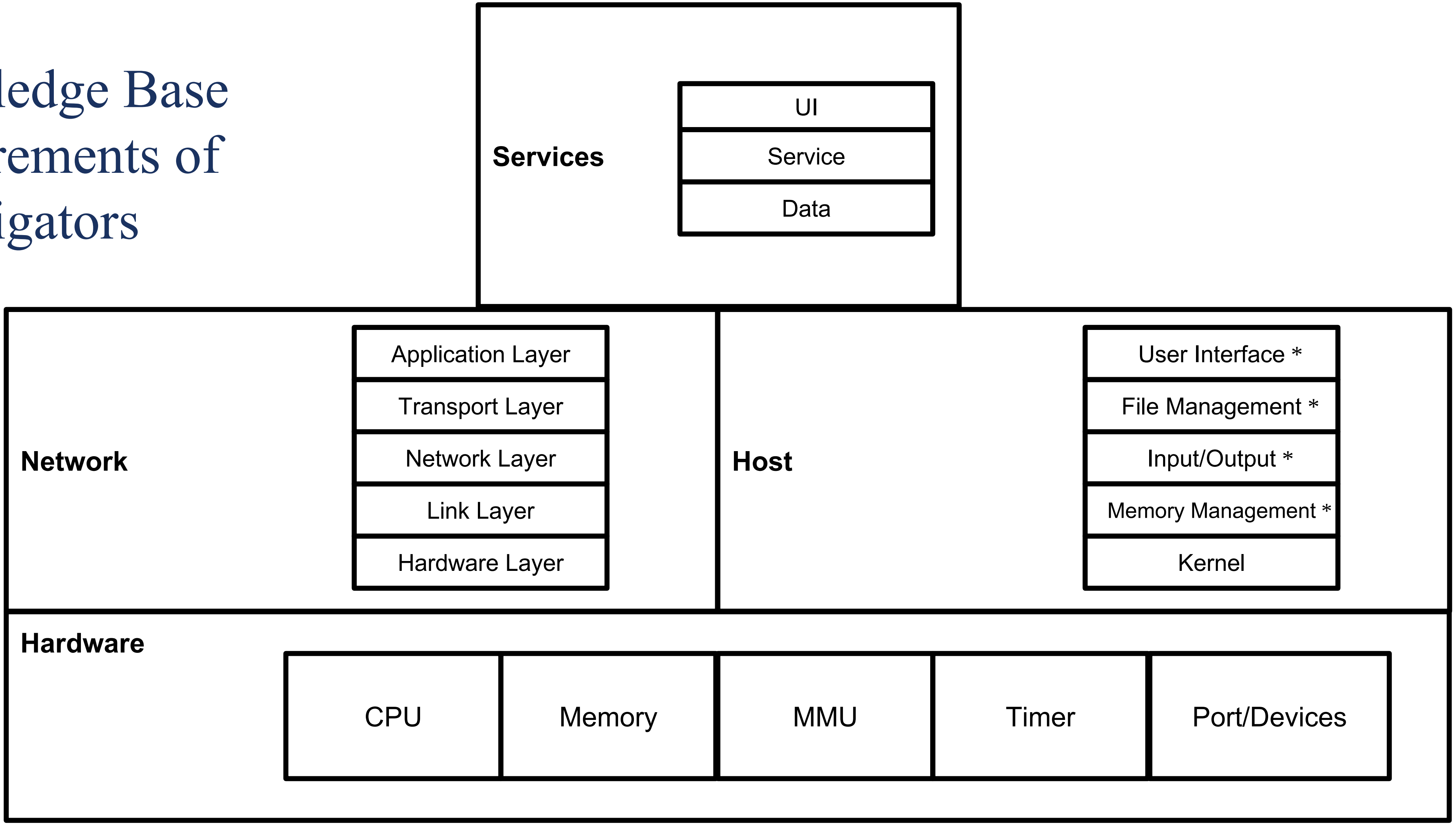
Yes	**No**	**Maybe** 8
0.63661017	0.46627119	0.55711864

Conclusion

- Digital forensic analysis is need of a reliable method
- Benefit from structure of mathematics, statistics & probability
  - Computer science research can assist the digital forensics community

Backup Slides

Knowledge Base Requirements of Investigators



Services

UI

Service

Data

Network

Application Layer

Transport Layer

Network Layer

Link Layer

Hardware Layer

Host

User Interface *

File Management *

Input/Output *

Memory Management *

Kernel

Hardware

CPU

Memory

MMU

Timer

Port/Devices

Degree Distribution

- Degree of a node in a graph is the number of connections it has to other nodes
- Degree distribution is the probability distribution of these degrees over the graph
- Degree distribution P(k) of a graph is then defined to be the fraction of nodes in the graph with degree k
- If there are n nodes in total in a graph and n_k of them have degree k, we have $P(k)=n_k/n$

Probabilistic Mass Function

- A function that gives the probability that a discrete random variable is exactly equal to some value
- Primary means of defining a discrete distribution

Bayesian Network

- Probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph
- Edges represent conditional dependencies
- Nodes that are not connected represent variables that are conditionally independent of each other
- Each node is associated with a probability function that takes a set of values for the node’s parent variables and gives the probability of variables represented by the node
- Attempt to alleviate the subjectivity in assigning prior probabilities through the probabilistic mass function

Normalizing Constant

- Reduce any probability function to a probability density function with total probability of one
- A constant by which an everywhere non-negative function must be multiplied so the area under its graph is 1