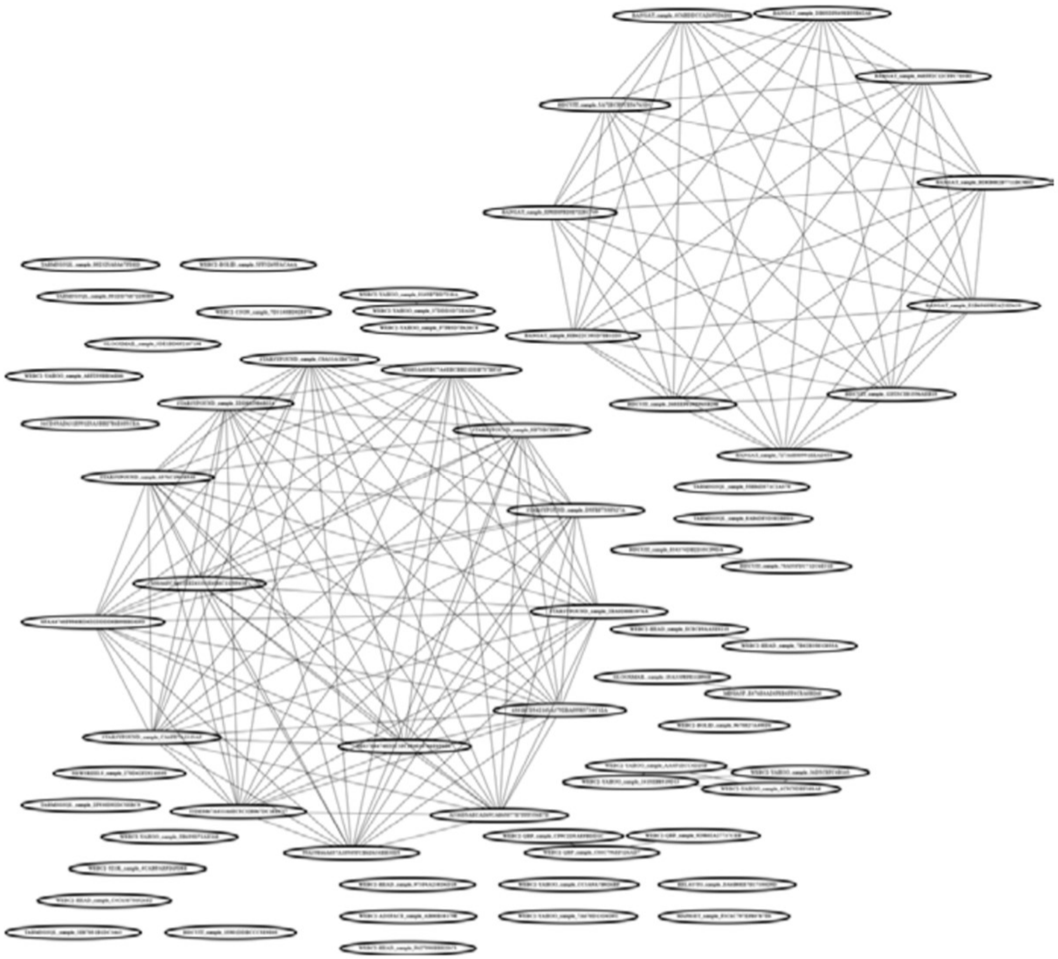
Chapter 4: Identifying attack campaigns using malware networks

DATA SCIENCE IN SECURITY

Introduction

- Malware network analysis
  - can turn malware datasets into valuable threat intelligence revealing
    - adversarial attack campaigns
    - common malware tactics
    - sources of malware samples.
  - consists of analyzing connections between groups of malware samples
    - embedded IP addresses
    - Hostnames
    - strings of print-able characters
    - Graphics
    - or similar

Introduction



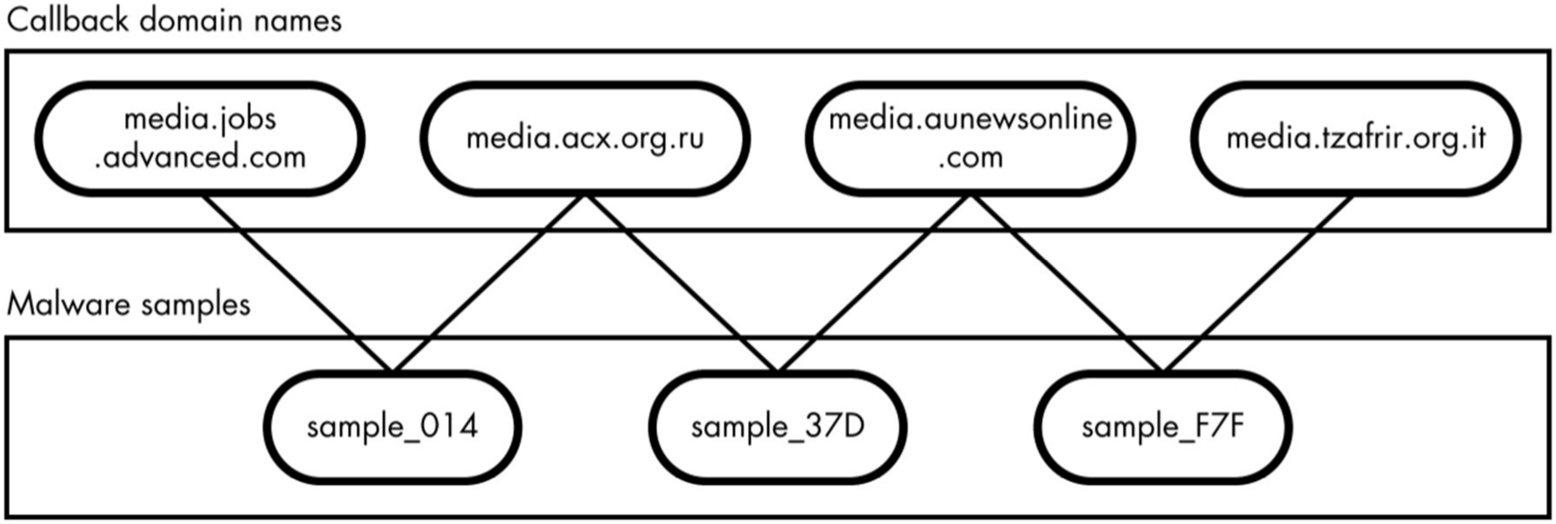
Connected samples “call back” to the same hostnames and IP addresses

Nodes and edges

- nodes and edges in a network can represent almost anything
  - we care about the structure of the interconnections
    - can reveal telling details about malware
  - we can treat
    - each individual malware file as a node
    - relationships of interest as an edge.
      - such as shared code or network behavior
    - Similar malware files cluster together
  - Also, we can treat
    - both malware samples and attributes as nodes
    - For example
      - callback IP addresses have nodes
      - malware samples have nodes.
      - malware samples nodes connect to corresponding IP address nodes

Bipartite networks

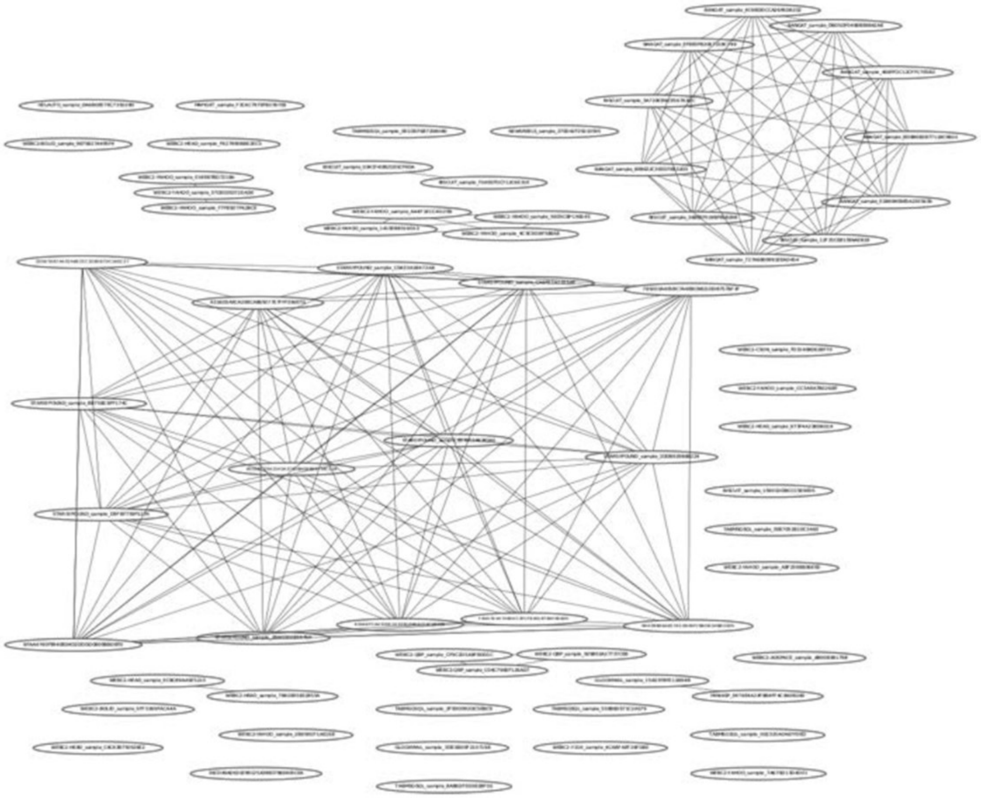
- A network whose
  - nodes can be divided into two partitions
  - neither partition contains internal connections
- can be used to show shared attributes between malware samples



Bipartite networks

- bipartite network projection
  - a simpler version of a bipartite network
  - link nodes in one partition of the network if they have nodes in the other partition in common

Bipartite networks



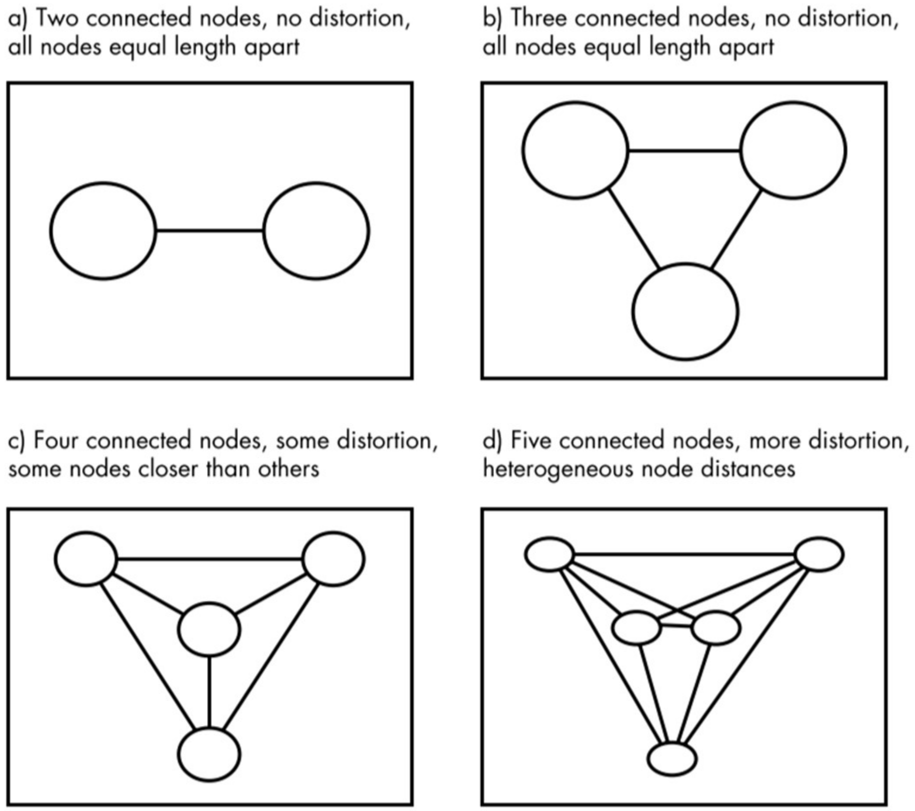
- projected network of the shared-callback servers of the entire Chinese APT1 dataset
- nodes are malware samples
- we can begin to see the overall “social network” of these malware samples

Visualizing Malware Networks

- the major challenge in network visualization is network layout
  - the process of deciding where to render each node within a 2D or 3D coordinate space
  - ideal way is to place them such that their visual distance from one another is proportional to the shortest-path distance between them in the network.
    - E.g.
      - nodes that are two hops away from one another might be about two inches away from one another
      - nodes that are three hops away might be about three inches apart
    - Doing this allows us to visualize clusters of similar nodes accurately to their actual relationship

Visualizing Malware Networks

- The Distortion Problem



a) Two connected nodes, no distortion, all nodes equal length apart

b) Three connected nodes, no distortion, all nodes equal length apart

c) Four connected nodes, some distortion, some nodes closer than others

d) Five connected nodes, more distortion, heterogeneous node distances

Figure 4-4: Perfect network layout is usually impossible on real-world malware networks. Simple cases like (a) and (b) allow us to lay out all nodes equidistantly. However, (c) adds distortion (the edges are no longer all equal length), and (d) shows even more distortion.

Visualizing Malware Networks

- Force-Directed Algorithms
  - are based on physical simulations of spring-like forces as well as magnetism
  - often leads to good node positioning

Demo

- Building networks with NetworkX
  - Adding nodes and edges
  - Adding Attributes
  - Saving Networks to Disk
- Network Visualization with Graphviz
  - Using Parameters to Adjust Networks
  - The GraphViz Command Line Tools
    - fdp
    - sfdp
    - neato

Demo

- Adding Visual Attributes to Nodes and Edges
  - Edge Width
  - Node and Edge Color
  - Node Shape
  - Text Labels