Recorded Future®

WHITE PAPER

The Security Intelligence Graph

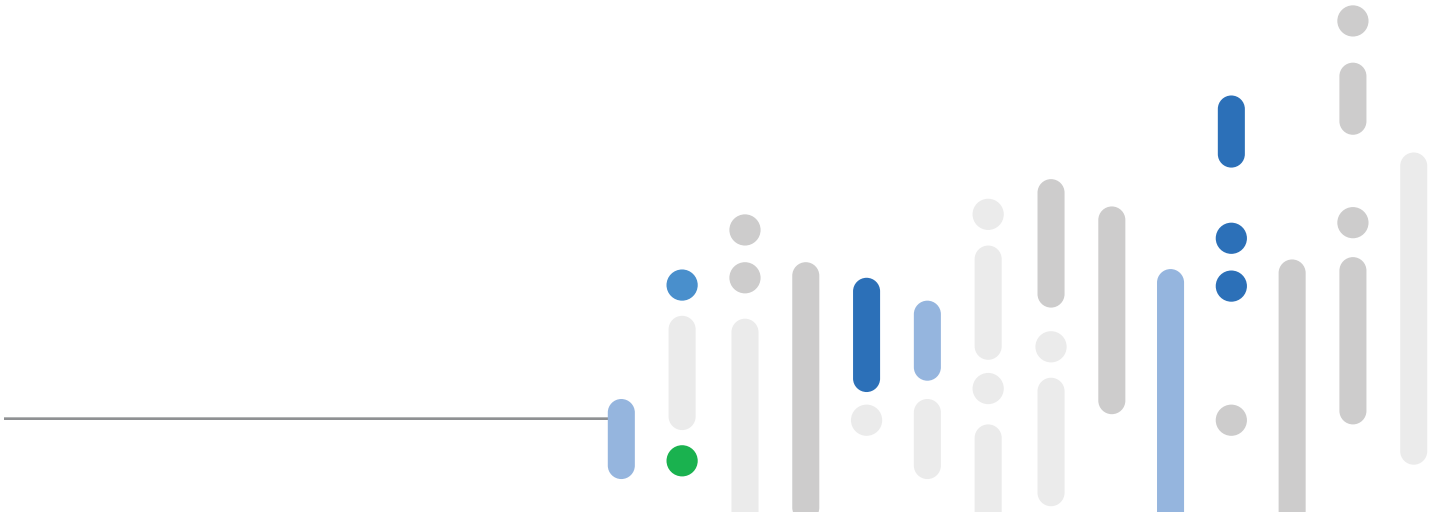

Table of Contents

Recorded Future captures all information gathered from the internet for over a decade and makes it available for analysis — we call this the Security Intelligence Graph. This paper describes the different components of the Security Intelligence Graph and how it is used to guide and drive analytic processes for both human analysts and algorithms. Our goal is not to demonstrate how to work with the Recorded Future product in particular, but to explain the underlying data model and philosophy of our system.

Recorded Future®

Introduction: Building a Digital Twin of the World

In 2009, Recorded Future was founded with the ambition to organize everything published on the internet and make it available for analysis, in real time. Our initial business focus was on government and military intelligence. That focus has since expanded to cyber threat intelligence, while maintaining the broader ambition of modelling all relevant security information available on the internet.

Just as many industrial companies today are creating "digital twins" of their products, we aim to build a digital twin of the world, representing all entities and events that are talked about on the internet — with a particular focus on threat intelligence. The Security Intelligence Graph is that representation of the world, and our goal is to make this information available at the fingertips of all security analysts to help them work faster and better.

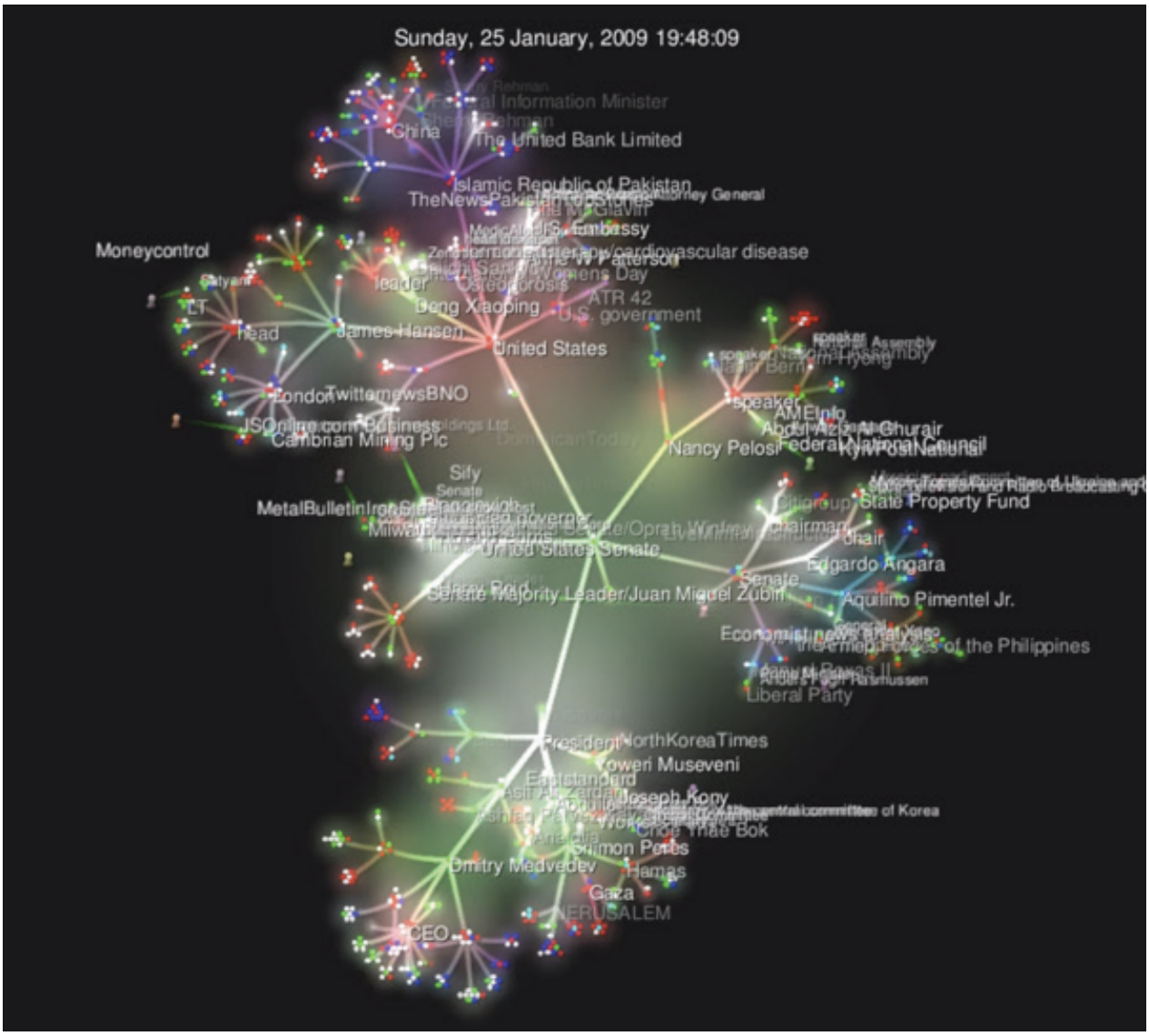

Figure 1: A very early (2010) visualization of the Security Intelligence Graph, in this case showing discussions on the internet related to Barack Obama. (See related video)

Our first step toward bringing our vision to life was manifested in our initial patent application in 2007, which outlined how to build a model of the world by harvesting all of the information published on the internet and organizing it for analysis.

A graph is an abstract mathematical object, consisting of nodes (representing entities, or "nouns" of the world) and edges (representing relationships between these entities). Graphs can be represented using different technologies, including both relational and graph databases, but the important thing is that graph algorithms operate on the abstract notion of nodes and edges. These algorithms allow us to find the shortest path between two cities, identify key influencers in a social network, or calculate the risk level of an internet domain name, to name a few examples.

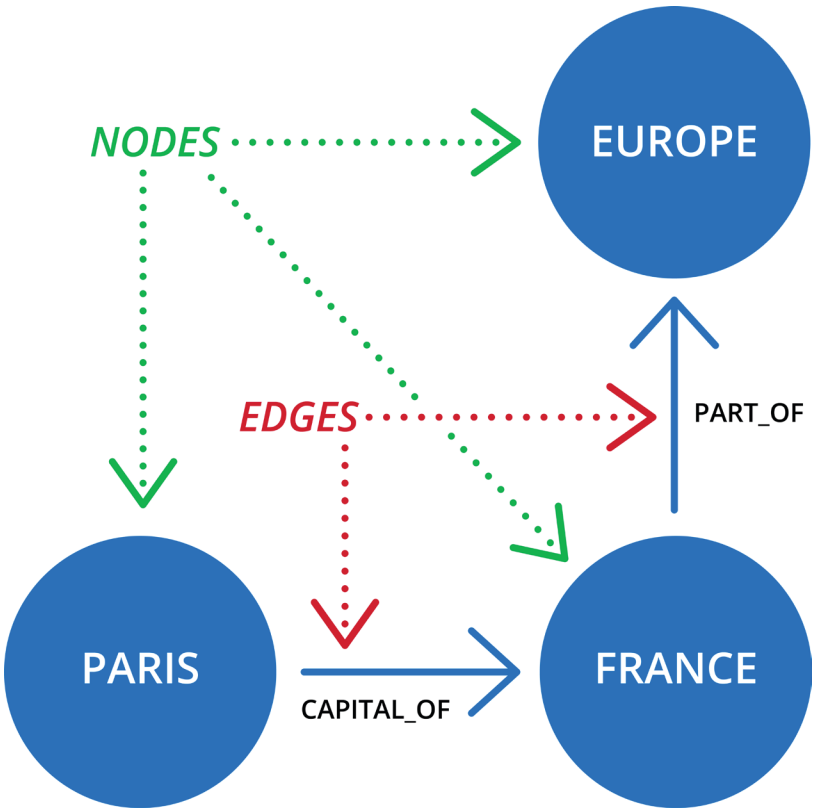

Figure 2: A small graph consisting of three nodes and two edges.

Recorded Future's co-founders come from a computer science and artificial intelligence background, so it was natural for us to organize information in a security intelligence graph made up of two parts: an ontology graph and an event graph.

Our ontology graph is used to represent slower-changing information about the world with a high degree of reliability. In the ontology graph:

- Nodes represent entities (both real-world entities like companies, people, places, products, and organizations, and abstract entities such as IP addresses, domain names, malware, attack vectors, and vulnerabilities).
- Edges between the nodes represent different relationships between entities (e.g., ownership, industry categories, geographical hierarchies, and technical dependencies

The other part of the Security Intelligence Graph is the event graph, which represents fast-changing and evolving information gathered from the internet. In the event graph:

- Nodes represent references (real-world events like cyberattacks and military maneuvers, as well as entity relationships). Other nodes represent clusters of such references.
- Edges between the nodes represent the grouping of references into events or clusters.

Finally, the two parts of the graph are connected by edges representing the relationships between events and the entities related to them. For example, a cyberattack event node can be related to an attacker and target entity node.

Nodes and edges in the Security Intelligence Graph can also have attributes. For reference nodes, these attributes describe metadata such as the original source, the media type, and the publication and event times[1] of a reference. References also have attributes computed from the text itself, such as sentiment scores. The entity nodes can have a large number of attributes, including Recorded Future risk scores and entity type-specific attributes such as the birth date of a "person" entity or the population of a "city" entity.

A security analyst can use these attributes to quickly select relevant subsections of the Security Intelligence Graph (e.g., by searching for events mentioned only on social media sources or company entities belonging only to a certain industry category).

[1] The event time for a reference is calculated based on the reference publication time and temporal expressions in the text, such as "next year" or "three days ago."

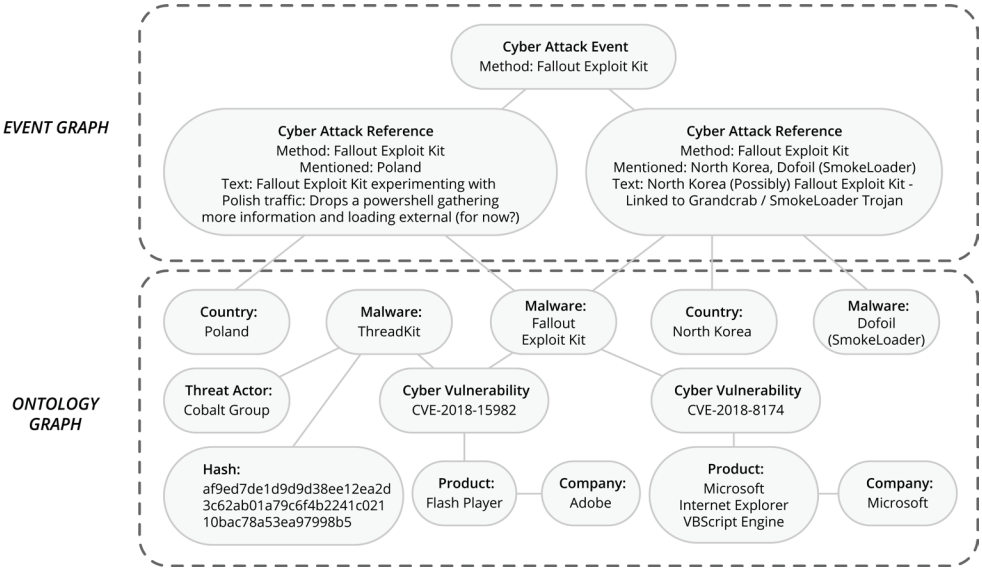

Figure 3: The Security Intelligence Graph and its two components, the event graph and the ontology graph.

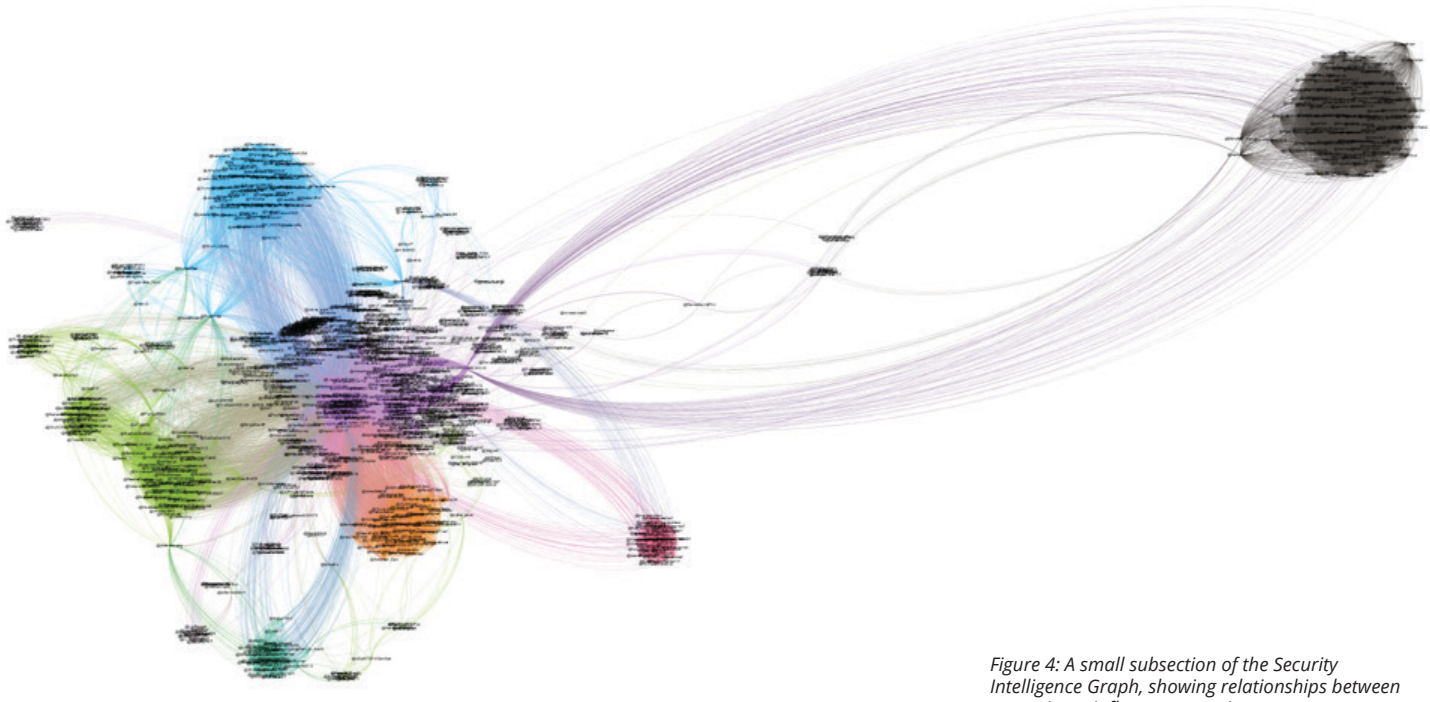

Figure 4: A small subsection of the Security Intelligence Graph, showing relationships between actors in an influence operation.

Navigating the Security Intelligence Graph

The Security Intelligence Graph allows both human analysts and algorithms to seamlessly pivot through complex relationships (e.g., when working with the Diamond Model of Intrusion Analysis). Figure 4 shows an example starting with a file hash and finding a related file, the vulnerability that file is related to, malware that exploits that vulnerability, a threat actor utilizing that malware, and in the end, a government organization associated with that threat actor.

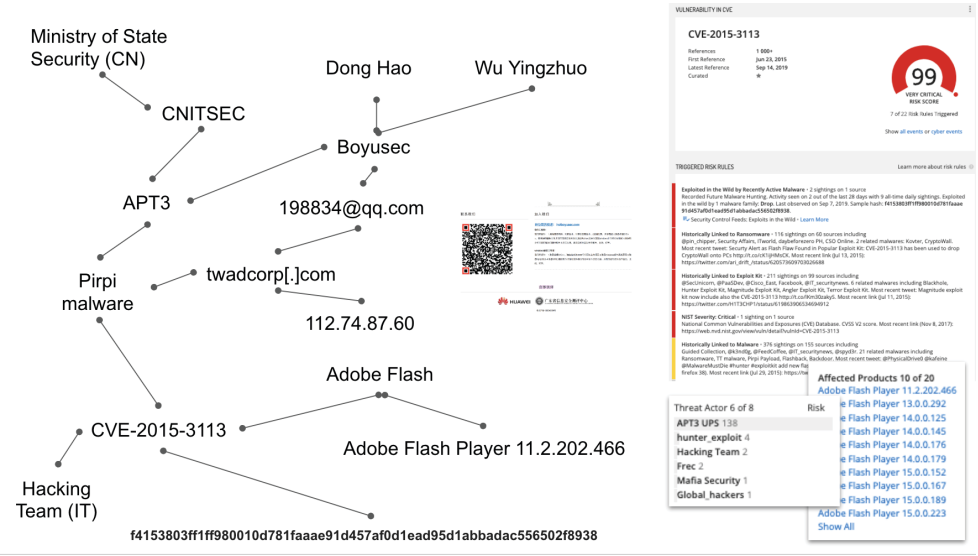

Figure 5: A subsection of the Security Intelligence Graph showing how it allows pivoting between hashes, files, vulnerabilities, malware, and threat actors.

Having all information readily available in the Security Intelligence Graph offloads a tremendous amount of work from analyst teams. It could take an organization thousands of man hours to build out a fraction of what is now readily available, and that time can instead be spent on analysis. By adding their own analyst notes, the team can even connect their own findings to the Security Intelligence Graph.

Intelligence Is a Volume Game

Over time, our Security Intelligence Graph has grown, and today consists of more than 63 billion reference nodes and more than four billion entity nodes, all connected by hundreds of billions of edges. On a daily basis, we typically add more than 40 million new reference nodes and three million new entity nodes to the graph. Computed attributes such as risk scores are all updated in real time for more than one billion entity nodes. The Security Intelligence Graph is comparable in size to the Google Knowledge Graph, but unlike that graph, the Security Intelligence Graph is updated in real time and focused on security intelligence.

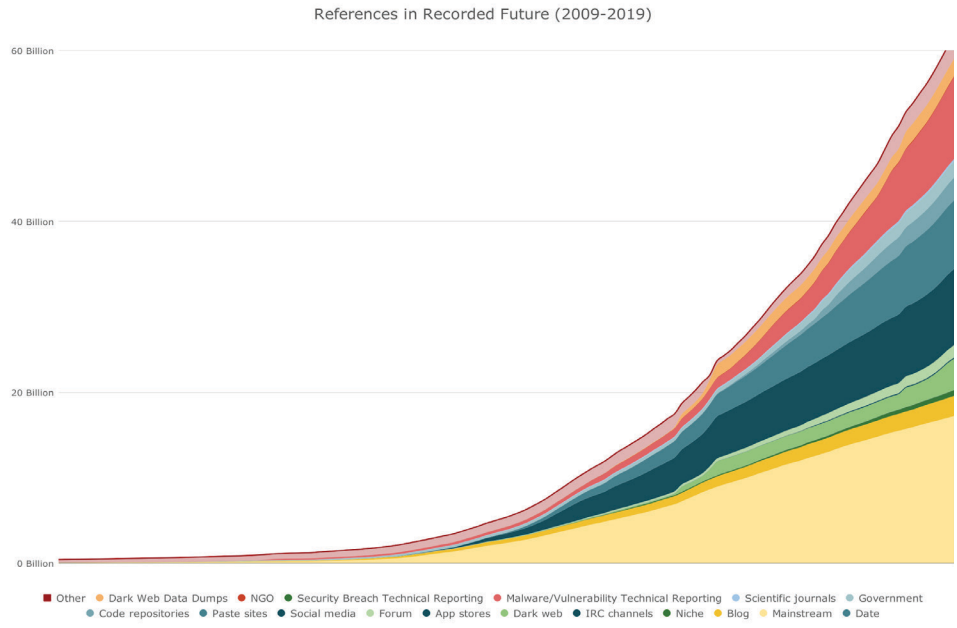

Figure 6: Total number of reference nodes over time, colored by media type.

Entity Type	Count
AS Number	85,306
Malware Families	138,626
Cyber Vulnerability	300,149
Technology	2,966,283
Company	23,755,650
File Name	98,590,378
IP Address	196,167,033
Internet Domain Name	291,637,650
URL	565,951,070
Hash	681,154,212

Figure 7: Number of entity nodes of different types.

Background: Graph Theory

In 1736, the Swiss mathematician Euler used a graphical representation to solve the problem of deciding if the citizens of Königsberg could take a walk that crosses all seven bridges in the city exactly once (turns out, they cannot).

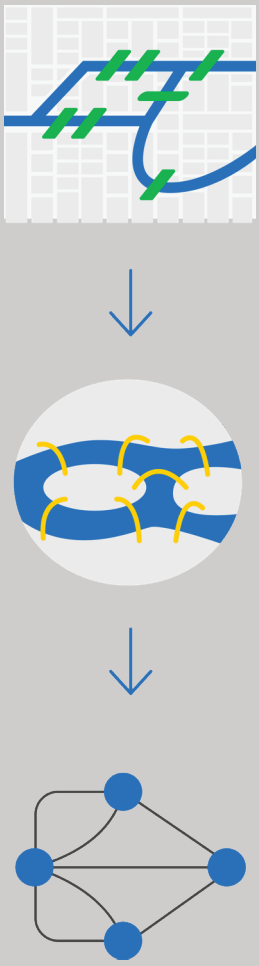

Figure 8

The graphical abstraction invented by Euler allowed him to both answer the burning question posed by the Königsberg citizens and prove general properties of graphs. Since then, graph theory has become a key method of describing complex relationships in a way that allows for algorithmic analysis.

Building the Security Intelligence Graph

The Security Intelligence Graph is constructed and updated from a number of sources. All text sources that are harvested by Recorded Future are analyzed using natural language processing (NLP) to extract entities, events, and temporal information. This information is used to create new, or update existing, entity nodes in the graph and to create new event reference nodes and edges between the entity and reference nodes. Technical sources are also used to create entity nodes, update their attributes, and sometimes create new reference nodes.

Ontological data is used to update the ontology graph with information about relationships between geographical entities, person entities, and which company entities they have a role in. Information from the NLP analysis of text is sometimes also used to update ontological relationships. Recorded Future uses this analytics machinery to compute risk scores for entities, which are stored as attributes of the entity nodes in the graph.

Insikt Group Analyst Notes are created using the same NLP process as external text resources, but can also contain explicit relationships between entities that are represented in the graph. Anonymized usage data from the Recorded Future product is also used to set attributes on entities. Finally, even though most work on building and updating the Security Intelligence Graph is done by algorithms, there is always a need for human curation to correct mistakes or add information that is not readily available in machine-readable form. Recorded Future's customer success and data science teams have a sophisticated set of tools to perform such manual curation work.

All Recorded Future clients also contribute to the evolution of the Security Intelligence Graph by regularly proposing new sources and pointing out missing or incorrect data. This drives both the automated collection and the manual curation work that is required to ensure the Security Intelligence Graph stays up to date and remains as comprehensive as possible. Therefore, this task is shared between Recorded Future algorithms and employees, and the entire Recorded Future user community.

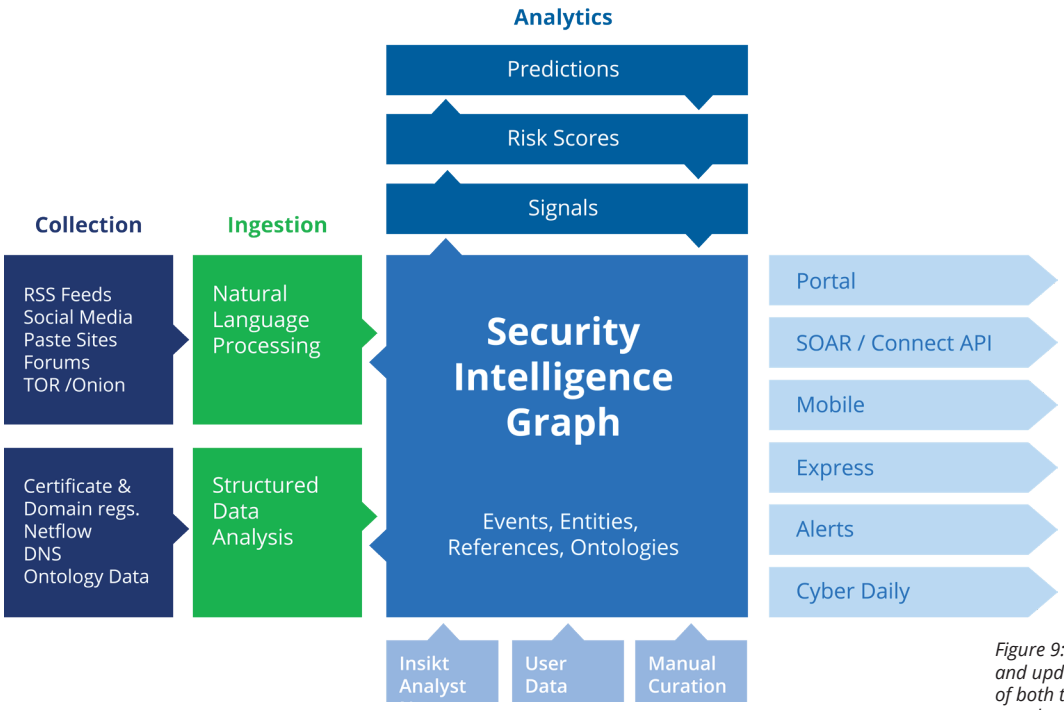

Figure 9: The Security Intelligence Graph is constructed and updated using open and closed source collection of both text-based resources and technical data, together with algorithmically computed scores, Insikt Group Analyst Notes, usage data from Recorded Future, and manual curation.

Computing Risk Scores With the Security Intelligence Graph

As aforementioned, we use the Security Intelligence Graph for different analytic purposes. One important application is the calculation of risk scores for entities, where we calculate an aggregate value for the risk associated with, for example, an internet domain name. Risk scores allow analysts and algorithms using our APIs to quickly make decisions on blocking a domain, for instance — knowing that the score takes into account all known critical information related to the domain. We use a rule-based approach to compute aggregate risk values for entities such as IP addresses, hashes, domains, and vulnerabilities.

Our third-party risk score for company entities relies the ontology graph heavily, using links between digital assets such as domains, IP address ranges, and ASNs and a company owning them. Other edges in the graph represent what technologies the company uses, which ones are its subsidiaries, etc. The reference graph represents what kinds of events the company has been involved in, and what vulnerabilities are currently being exploited in the wild that are affecting products in use by the company. All of these relationships are used to compute a risk score per company, and every time the relevant parts of the graph change, the risk score is recalculated and the corresponding node attributes are updated.

The Recorded Future APIs

Recorded Future clients can access the Security Intelligence Graph programmatically through different APIs:

- The Connect API does not expose the Security Intelligence Graph as a set of nodes and edges, but instead provides the ability to look up entity nodes and their attribute values, as well as aggregated information such as reference statistics, co-occurring entities, and risk metrics (triggered risk rules and risk scores). Entities can be searched by name or looked up by using various filters such as risk score ranges, risk rules triggers, and entity ontology links.
- The security orchestration, automation and response (SOAR) API enables very high-frequency lookups of risk information for IOCs and more in-depth use case-based risk assessments to be used in integrations with external orchestration systems.
- Recorded Future Fusion allows for programmatic use of the Security Intelligence Graph for correlation and enrichment of customer data.

The goal of Recorded Future's APIs is to expose the information contained in the Security Intelligence Graph to our clients in a role-appropriate manner. Different security roles within an organization require access to different views of Security Intelligence Graph. For example, the vulnerability management team needs access to the vulnerability intelligence contained within the Connect API to understand how they should prioritize patching based on risk score. On the other hand, security leadership would be more interested in what vulnerabilities are being exploited by groups targeting the organization, so the leadership team would need more specific and focused queries available through the Recorded Future Fusion API.

Each API can be customized to provide not only a role-specific view of the Security Intelligence Graph, but also a platform-specific view. The Connect API can provide high-volume intelligence that can be used for correlation purposes in a security information and event manager (SIEM), and it can also provide focused client-specific alerts that can be ingested through a ticketing system using the Alert API endpoint, which is part of the Connect API. Feeding the Alert API into a ticketing system allows security teams to take more immediate action on high-priority issues, such as a new typosquat domain registrations or organizational infrastructure being used by cybercriminals for command and control purposes.

Many organizations have unique intelligence requirements that cannot be easily met by a broad set of indicators like those provided through the Recorded Future Connect API. In those cases, the Recorded Future Fusion API can provide access to a more narrow view of the Security Intelligence Graph. Threat intelligence analysts can use the Recorded Future Fusion API to track specific queries over time, or merge intelligence present in the Security Intelligence Graph with third-party intelligence to provide a more specific picture of a topic or entity.

Ultimately, the Recorded Future APIs allow clients to integrate the information available in the Security Intelligence Graph directly into their workflows in a programmatic fashion that is specific to either the role or tool being used by the person in that role.

Future Evolution of the Security Intelligence Graph

The Security Intelligence Graph continues to grow. Millions of nodes and edges are added every day — in reflection of the living, breathing heartbeat of the internet. In addition to this organic growth, we constantly add new sources, new entity and event types, and new analytics. At Recorded Future, our goal is for the Security Intelligence Graph to always be the most comprehensive source for security analysts, with the ambition to protect their organizations or nations from present or future threats. Our belief is that threat analyst centaurs — the seamless combination of algorithms and humans — is the only way to achieve this goal, and the Security Intelligence Graph is the fabric enabling that collaboration.

www.recordedfuture.com

@RecordedFuture

About Recorded Future

Recorded Future arms security teams with the only complete threat intelligence solution powered by patented machine learning to lower risk. Our technology automatically collects and analyzes information from an unrivaled breadth of sources and provides invaluable context in real time and packaged for human analysis or integration with security technologies.