Selected Papers of #AoIR2021:
The 22nd Annual Conference of the
Association of Internet Researchers
Virtual Event / 13-16 Oct 2021

THE SEER AND THE SEEN: A SURVEY OF TOPICS FOUND IN PALANTIR PATENTS

Andrew Iliadis
Temple University

Amelia Acker
University of Texas at Austin

Introduction

Palantir is one of the most secretive technology firms in the US. The company supplies information technology solutions to government agencies, humanitarian organizations, and corporations, focusing specifically on data integration and surveillance services. Several qualitative studies have examined use of Palantir's products in field operations by police agencies through ethnography and legal case studies (Brayne, 2020; Ferguson, 2017) or have conducted critical and rhetorical analyses of Palantir's marketing, reports, and the public-facing associated literature describing Palantir's software products and services (Munn, 2017; Knight and Gekker, 2020).

To investigate Palantir's opaque technology practices, this paper presents findings from a computational topic modeling of a purposive sample (n=155) of Palantir's patents filed from 2006-2019 in the US, Germany, Australia, UK, and EU, along with a description of key patent topics and themes. This approach follows the spate of recent literature that uses patents as primary data for researching opaque information technology firms.

In recent years, press reporting has covered Palantir's links to the US National Security Agency's (NSA) surveillance operations through the Edward Snowden whistleblowing revelations, accused Palantir of human rights violations "targeting parents and caregivers of unaccompanied migrant children" according to Amnesty International, and reproached Palantir's unethical approaches to responsible corporate conduct. Yet, Palantir recently received a valuation totaling $20 billion USD and the number of academic studies about Palantir's organization and software can be counted on one hand. This article contributes to this scholarship by providing firsthand, primary source documentation of Palantir's surveillance platform, explaining how the company imagines and describes its technical capabilities.

Iliadis, A. and Acker, A. (2021, October). *The Seer and the Seen: A Survey of Topics Found in Palantir Patents*. Paper presented at AoIR 2021: The 22nd Annual Conference of the Association of Internet Researchers. Virtual Event: AoIR. Retrieved from http://spir.aoir.org.

Methods

For this study, we scraped all Palantir's "ontology" patents (as of 08/25/20) from Google Patents. This produced a purposive sample (n=155) of Palantir patents, consisting of 5197 pages, over 2.5 million words, and over 18.5 million characters. We then prepared the dataset for processing by stripping all the metadata and special features, converting formats, compressing, and collating the patents together. Topic modeling was performed using a bag-of-words model and Latent Dirichlet Allocation (LDA). We collaboratively and iteratively reviewed the results and identified 20 topics that were modeled from the output containing 20 most frequent words related to each topic. We then developed 3 overarching themes that emerged from these 20 topics.

Findings

Some of the more interesting patents in our dataset include those with titles describing data integration, context-building processes, entity, property, and relationship identification, and threat detection. Table 1 below is a small sample of patent titles extracted from our corpus.

Table 1. Sample list of Palantir ontology patent titles.

- Federated search of multiple sources with conflict resolution
- Systems and methods for visual definition of data associations
- System and method for detecting confidential information emails
- System and method for evaluating network threats and usage
- Systems and user interfaces for dynamic and interactive access of, investigation of, and analysis of data objects stored in one or more databases
- Systems and user interfaces for holistic, data-driven investigation of bad actor behavior based on clustering and scoring of related data
- Relationship visualizations
- Automated database analysis to detect malfeasance
- Systems and user interfaces for dynamic and interactive investigation of bad actor behavior based on automatic clustering of related data in various data structures
- System and method for sharing investigation results

Along with titles, the corpus included metadata for patent ID codes, assignee names, inventor's names, priority dates, filing and creation dates, publication dates, result links, and representative figure links, among other common structured information found in technology patents. Among the 155 patents only 51 have been granted, indicating that at the time of collection roughly a third of Palantir's ontology patents had been granted by the patent and trademark offices of their respective countries. This does not mean that a large majority of these patents will fail to become granted, as the longest observable time between priority date (first date) and grant date (final date) was 12 years, from 11/20/2006 to 8/28/2018, and many patents among the corpus were given priority dates only within the last few years. Among the countries and regions in which the patents were filed, the breakdown was 31 from the European Patent Office, 4 from Germany, 1 from Australia, 1 from the UK, 1 from the Netherlands, and 117 from the US, clearly showing that Palantir files most of their ontology patents domestically. Among proper names, we distilled platform companies' names from the corpus. These included Amazon, Apple, Facebook, Google, Instagram, LinkedIn, Microsoft, and

Twitter. Google and Microsoft were mentioned much more overall than the other companies in the Palantir patents, usually in the context of integrating data from the services that they offer. The data here shows that Palantir envisions its products integrating data from the products and services of platforms. Table 2 below shows the custom proper name keywords and most common proximal word frequencies.

Table 2. Custom proper name keywords and most common proximal word frequencies.

Amazon		Apple		Facebook		Google		Instagram		LinkenIn		Microsoft		Twitter	
repository	12	apple	14	social	4	google	225	platform	5	social	4	microsoft	226	twitter	18
data	11	attribute	9	consumer	3	email	66	medium	4	platform	3	page	100	followers	14
storage	7	stock	9	platform	3	page	55	social	4	consumer	3	printed	84	visualizing	14
amazon	6	data	8	medium	2	trader	45	consumer	2	medium	2	word	83	tools	14
service	6	example	7	facebook	2	retrieved	39	instagram	2	facebook	2	document	63	great	14
received	5	associated	6	linkedin	2	analytics	38	network	2	linkedin	2	retrieved	61	article	13
device	5	product	6	location	2	maps	37	among	1	location	2	format	39	amnet	11
queue	5	trade	5	browser	1	internet	36	group	1	network	2	example	37	html	10
simple	5	generating	4	posted	1	printed	36	related	1	group	1	connector	36	posted	9
contains	5	trader	3	information	1	trade	36	behavior	1	information	1	glue	36	postedaug	8
source	5	cash	3	similar	1	earth	30	online	1	similar	1	nector	35	platform	7
interface	5	advertisement	3	square	1	example	30	available	1	square	1	connecting	34	data	6
programming	5	spicy	3	twitter	1	using	28	linkedin	1	twitter	1	shape	34	medium	6
code	5	browser	3	provide	1	data	24	facebook	1	provide	1	tool	32	modality	6
application	4	price	3	entity	1	search	23	others	1	entity	1	template	29	social	6
message	3	include	3	group	1	deep	21	networking	1	example	1	data	29	location	5
computing	3	bucket	3	behavior	1	manual	21	like	1	related	1	system	27	example	5
file	3	particular	3	related	1	drew	21	provider	1	behavior	1	jose	27	dated	5
like	2	value	3	online	1	missing	21	interaction		online	1	aspx	27	flow	4
tory	2	large	2	available	1	dive	21	known		available	1	application	27	design	4

Our main analysis object was the topic modeling, which included the topics, associated keywords for each topic, and our own manually chosen examples taken from the data in the form of excerpts. The two sets of topics, keywords, and examples are presented in Tables 3 and 4.

Table 3. Topic modeling for topics 1-10.

Topic	Most Frequent Words	Examples
Topic 1. Identifying data matches, associations, and their relationships	'associated', 'match', 'subset', 'collection', 'matching', 'query', 'plurality', 'review', 'set', 'based', 'information', 'action', 'ni', 'data', 'field', 'resource', 'message', 'criterion', 'result', 'search'	"The definition Name:Last, Name:First specifies that matching input data values map to components named "Last" and "First" of the Name property."
Topic 2. Memory, processing, and storage devices	'disk', 'execution', 'stored', 'executed', 'main', 'bus', 'example', 'cause', 'configured', 'software', 'data', 'hardware', 'module', 'device', 'memory', 'storage', 'processor', 'medium', 'instruction', 'computer'	"Accordingly, a "machine-readable medium" refers to a single storage apparatus or devices, as well as "cloud-based" storage systems or storage networks."
Topic 3. Selecting customizable data objects in graphs with nodes and edges	'receiving', 'command', 'using', 'new', 'example', 'embodiment', 'interface', 'cursor', 'selected', 'device', 'parser', 'type', 'information', 'control', 'selection', 'display', 'graph', 'node', 'user', 'input'	"Another type of user input device is cursor control, such as a mouse, a trackball, or cursor direction keys for communicating direction information and command."
Topic 4. Modeling and linking ontologies, entities, objects, relationships, and properties related to people, events, and things	'represent', 'model', 'identifier', 'document', 'database', 'ontology', 'attribute', 'particular', 'based', 'embodiment', 'associated', 'example', 'relationship', 'link', 'person', 'event', 'property', 'type', 'data', 'object'	"The curated ontology may thereby be delivered to the presentation module in order to generate presentation of the data object based on the curated ontology."
Topic 5. Segmentation and partitioning objects using sets, claims, time series, and values	'data', 'attribute', 'subset', 'parameter', 'segment', 'widget', 'time', 'associated', 'concept', 'respective', 'comprises', 'plurality', 'method', 'determining', 'based', 'set', 'display', 'claim', 'value', 'second'	"Rendering one or more graphical data display widgets according to each display order value, which widget type and the widget configuration values."
Topic 6. Provenance, receipts, and verification of claims, descriptions, disclosures	'include', 'specific', 'understood', 'certain', 'description', 'intended', 'claim', 'used', 'present', 'scope', 'element', 'invention', 'disclosure', 'feature', 'term', 'official', 'embodiment', 'communication', 'dated', 'appl'	"By generating and storing metadata in association with externally-generated datasets, provenance and lineage of the external builds can be maintained."
Topic 7. Logical rule-based scoring and analysis with alerts	'rule', 'particular', 'list', 'scoring', 'strategy', 'according', 'based', 'user', 'various', 'associated', 'analysis', 'generated', 'embodiment', 'engine', 'example', 'score', 'data', 'alert', 'analyst', 'cluster'	"The cluster/rules engine is used in conjunction with data stored in database, with the data optionally being stored in data tables, as described above."
Topic 8. Communication and notification with messages	'message', 'connection', 'corresponding', 'metric', 'computer', 'type', 'local', 'link', 'internet', 'interface', 'sensor', 'task', 'machine', 'example', 'maintenance', 'fault', 'data', 'communication', 'log', 'network'	"A notification is displayed in response to determining that the one or more validation rules exclude at least one constraint from the set of constraints."
Topic 9. Visualizations with process flowcharts, figures, and diagrams	'process', 'exemplary', 'web', 'flowchart', 'official', 'block', 'figure', 'dated', 'diagram', 'disclosure', 'patent', 'according', 'communication', 'present', 'document', 'illustrates', 'example', 'embodiment', 'application', 'fig'	"Each block in the flowchart or block diagrams may represent a module, segment, or portion of instructions, which comprises one or more executable instructions."
Topic 10. Medical and healthcare fraud claims	'particular', 'medical', 'service', 'associated', 'claim', 'pharmacy', 'fraud', 'patent', 'date', 'summary', 'patient', 'healthcare', 'view', 'metric', 'description', 'ul', 'lead', 'provider', 'panel', 'filter'	"The healthcare fraud attack alert may display a risk/importance level, and/or "from" entity indicating the source of the healthcare fraud attack alert."

Table 4. Topic modeling for topics 11-20.

Topic 11. Client-side code, datasets, templates	'bar', 'set', 'used', 'instrument', 'dataset', 'code', 'period', 'function', 'different', 'selected', 'embodiment', 'chart', 'panel', 'series', 'template', 'user', 'example', 'value', 'query', 'time'	"The template may specify the configurations and/or settings for generating the customized user interface."
Topic 12. Levels of abstraction using layers, maps, graphs	'selected', 'information', 'feature', 'tile', 'various', 'object', 'graphical', 'interactive', 'level', 'view', 'layer', 'displayed', 'display', 'example', 'embodiment', 'data', 'access', 'map', 'interface', 'user'	"An interface may be provided between the overview layer and the underlying system or data source in order to provide a level of abstraction."
Topic 13. Server-side software operation, implementation, and performance	'implemented', 'electronic', 'software', 'windows', 'program', 'operating', 'perform', 'ca', 'described', 'include', 'embodiment', 'client', 'logic', 'mobile', 'technique', 'computer', 'component', 'server', 'computing', 'device'	"For example, the computing system may comprise a server system that accesses law enforcement data and provides user interface data to one or more users."
Topic 14. Schemas with sets, types, and properties	'include', 'map', 'transformation', 'ip', 'embodiment', 'database', 'definition', 'model', 'schema', 'component', 'information', 'value', 'set', 'address', 'example', 'ontology', 'source', 'type', 'property', 'data'	"Schemas define the structure of data sources for example, the names and other characteristics of tables, files, columns, fields, proper ties, and so forth."
Topic 15. Sandboxing, cloud, experimenting, running tests, decision making, analytics	'component', 'order', 'vector', 'update', 'disclosed', 'operation', 'version', 'replication', 'database', 'various', 'state', 'process', 'embodiment', 'performed', 'example', 'method', 'described', 'site', 'change', 'block'	"Recipient of the original master ontology may know to stop working on their version of the ontology in their sandbox."
Topic 16. Financial fraud detection	'embodiment', 'activity', 'analysis', 'trader', 'generation', 'financial', 'external', 'associated', 'trade', 'detection', 'account', 'email', 'strategy', 'example', 'related', 'tag', 'seed', 'cluster', 'item', 'data'	"A suspect stock trade for Google's stock may be identified. Associated with that trade may be a first trader who initiated the sell order."
Topic 17. Patents, references, literature, studies, and documentation	'conference', 'titled', 'pat', 'entirety', 'claim', 'sheet', 'provisional', 'systems', 'vol', 'publication', 'reference', 'incorporated', 'ser', 'patent', 'filed', 'bl', 'application', 'pp', 'et', 'al'	"Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days."
Topic 18. Geo-location, filtering/panes	'represent', 'step', 'geographic', 'filter', 'pane', 'number', 'include', 'performance', 'analysis', 'example', 'embodiment', 'data', 'consuming', 'category', 'associated', 'interaction', 'information', 'provisioning', 'location', 'entity'	"For example, the user may perform a query for one or more datasets such as finding call entries in cellular call databases associated with a particular geographic location."
Topic 19. Security, attack, fraud, kill chain, hacking	'plurality', 'module', 'processing', 'security', 'unit', 'attack', 'receiving', 'ontology', 'query', 'datasets', 'sharing', 'fraud', 'repository', 'database', 'comprising', 'claim', 'method', 'request', 'source', 'data'	"The attack data unit may be configured to receive a plurality of security attack data from one or more entities."
Topic 20. Representation and meaning, internal standardization and structure, metadata	'attribute', 'view', 'row', 'associated', 'set', 'format', 'user', 'stored', 'structure', 'embodiment', 'metadata', 'store', 'information', 'table', 'analysis', 'example', 'file', 'database', 'item', 'data'	"The alert display further includes a table of information associated with the one or more data items and associated metadata of the particular data cluster."

Finding 1: Labeling Human Traces and Sorting Actions

The first thread (Table 5) represents topics related to labeling human traces, sorting actions, and identifying normative flows of actions in information systems to flag fraud, alleged criminality, hacking, or unusual events. By labeling data with formal ontologies, knowledge of detection, prediction, and analysis can commence (i.e., beginnings of the data funnel).

Table 5. Finding 1: Labeling human traces and sorting actions.

- Topic 4. Modeling and linking ontologies, entities, objects, relationships, and properties related to people, events, and things
- Topic 5. Segmentation and partitioning objects using sets, claims, time series, and values
- Topic 10. Medical and healthcare fraud claims
- Topic 16. Financial fraud detection
- Topic 19. Security, attack, fraud, kill chain, hacking

Finding 2: Leveraging Ontologies, Semantic Data Structures for Integration

It is in the second thread where we can see themes related to the development of schemas, graphs, and ontologies for data integration in service of network ties and the visualizing of objects and the relationships between them. The topics in this second theme represent a higher level of abstraction from topics in the first theme, and instead the focus here is on second order meaning that emerges from observing big data from several heterogenous sources. These trends rely on compilation, and meaningful assembly in volume of databases from different domains. At this scale, compelling

evidence is found between these disparate domain entities, instead of labeling objects, events, and actions.

Table 6. Finding 2: Leveraging ontologies, semantic data structures for integration.

- Topic 1. Identifying data matches, associations, and their relationships
- Topic 3. Selecting customizable data objects in graphs with nodes and edges
- Topic 12. Levels of abstraction using layers, maps, graphs
- Topic 14. Schemas with sets, types, and properties
- Topic 20. Representation and meaning, internal standardization and structure, metadata

Finding 3: Data Work, Interpretation, Processing for Management, Analytics, Prediction

The last strand of topics (Table 7) reveals the software as a service items that Palantir provides to its customers, that is, the ability to make meaningful representations out of the information that users receive in the form of dashboards, visualizations, interfaces, documentation, communication, etc. These topics focus on how Palantir allows its users to manage data at scale using informative client or server-side systems to support prediction and decision analytics. Actionable items occur at this end stage of the data funnel, provided by data that have been semantically baked through the data work, interpretation, and processing for management, analytics, and prediction that occurs in Palantir's platform ecosystem that supports knowledge workers and data professionals.

Table 7. Finding 3: Data work, interpretation, processing for management, analytics, prediction.

- Topic 2. Memory, processing, and storage devices
- Topic 6. Provenance, receipts, and verification of claims, descriptions, disclosures
- Topic 7. Logical rule-based scoring and analysis with alerts
- Topic 8. Communication and notification with messages
- Topic 9. Visualizations with process flowcharts, figures, and diagrams
- Topic 11. Client-side code, datasets, templates
- Topic 13. Server-side software operation, implementation, and performance
- Topic 15. Sandboxing, cloud, experimenting, running tests, decision making, analytics
- Topic 18. Geo-location, filtering/panes

References

Brayne S (2020) *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford: Oxford University Press.

Ferguson AG (2017) *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. New York University Press.

Knight E and Gekker A (2020) Mapping interfacial regimes of control: Palantir's ICM in America's post-9/11 security technology infrastructures. *Surveillance and Society 18*(2): 231–243.

Munn L (2017) Seeing with software: Palantir and the regulation of life. *Studies in Control Societies 2*(1): 1–16.