




May 18th, 8:45 AM - 9:20 AM

## Exploring the Use of Graph Databases to Catalog Artifacts for Client Forensics

Rose Shumba  
shumba@usna.edu

Follow this and additional works at: <https://commons.erau.edu/adfsl>

 Part of the [Aviation Safety and Security Commons](#), [Computer Law Commons](#), [Defense and Security Studies Commons](#), [Forensic Science and Technology Commons](#), [Information Security Commons](#), [National Security Law Commons](#), [OS and Networks Commons](#), [Other Computer Sciences Commons](#), and the [Social Control, Law, Crime, and Deviance Commons](#)

---

### Scholarly Commons Citation

Shumba, Rose, "Exploring the Use of Graph Databases to Catalog Artifacts for Client Forensics" (2018). *Annual ADFSL Conference on Digital Forensics, Security and Law*. 5.  
<https://commons.erau.edu/adfsl/2018/presentations/5>

This Peer Reviewed Paper is brought to you for free and open access by the Conferences at Scholarly Commons. It has been accepted for inclusion in Annual ADFSL Conference on Digital Forensics, Security and Law by an authorized administrator of Scholarly Commons. For more information, please contact [commons@erau.edu](mailto:commons@erau.edu).

**EMBRY-RIDDLE**  
Aeronautical University™  
SCHOLARLY COMMONS

(c)ADFSL



# EXPLORING THE USE OF GRAPH DATABASES TO CATALOGUE ARTIFACTS FOR CLIENT FORENSICS

Rose Shumba  
US Naval Academy  
Cyber Science Department  
Annapolis, MD  
shumba@usna.edu

## ABSTRACT

Cloud computing has revolutionized the methods by which digital data is stored, processed, and transmitted. It is providing users with data storage and processing services, enabling access to resources through multiple devices. Although organizations continue to embrace the advantages of flexibility and scalability offered by cloud computing, insider threats are becoming a serious concern as cited by security researchers. Insiders can use authorized access to steal sensitive information, calling for the need for an investigation. This concept paper describes research in progress towards developing a Neo4j graph database tool to enhance client forensics. The tool, with a Python interface, allows for the location of evidential artifacts promptly. Initially, the database contains artifacts from existing research that can be used to prove usage. The ultimate goal is to create an Open Source collaborative environment for researchers and practitioners to add artifacts as we go along. The reasons for choosing a graph database are presented in the paper.

**Keywords:** cloud forensics, cloud storage services, client forensics, artifacts, cloud storage interaction

## 1. INTRODUCTION

Cloud computing in its various forms has become a staple paradigm for businesses, governments, and individuals in recent years, with Storage as a Service (StaaS), becoming increasingly popular (Top 10 Security Concerns, n.d; Six security risks of enterprises, n.d). Cloud storage services, such as Google Drive, Dropbox, and OneDrive, allow consumers to store, share, collaborate, synchronize, and edit data files via a range of devices, such as personal computers

and mobile devices (Faheen & Khan, 2014; Cloud Storage, n.d).

Even though organizations continue to embrace the advantages of flexibility, affordability, and scalability provided by cloud storage services, several security risks are prevalent. InfoWorld published twelve security threats organizations face when using cloud storage services (The dirty dozen, n.d). Among the twelve are insider threat and data breaches.

According to the Gartner Cloud Adoption and Risk report, 23.2% of security incidents

experienced by an organization are cloud related, with 93.5% of those being insider threats (Cloud Adoption and Risk, n.d). An insider can use authorized access to an organization's cloud storage-based services to misuse or steal sensitive or confidential company data (classified documents, intellectual property, trade secrets) (Narayan & Kaushik, 2016) (The state of cloud computing, n.d; Mills, 2012). Breaches involving trade secrets and intellectual property can be devastating. When these occur, organizations may incur fines or may face lawsuits or criminal charges (Narayan & Kaushik, 2016).

The exponential increase in the use of cloud storage-related services, the commensurate rise in security risks, and the growth in the level of threats posed by insiders has given rise to the need for better approaches and tools for cloud forensics, which in turn has brought to light several additional challenges.

The National Institute of Standards and Technology (NIST, 2014) and several researchers (Marturana, Me, & Tacconi, 2012; Quick & Choo, 2013; Quick, Martini, & Choo, 2014; Malik, Shashidhar, & Chen, 2015; Long, & Qing, 2015; Epifani, 2013) identified over 65 challenges associated with cloud forensics.

One of these challenges is the fact that investigators have limited access to physical servers to conduct server analysis. This presents the investigator with three options; to attempt to recover evidence from seized local devices known to have interacted with the cloud; to try and eavesdrop network traffic between local devices and the cloud network; to request a court in a foreign jurisdiction to seize evidence directly from a cloud server (Marturana, Me, & Tacconi, 2012). The latter brings additional legal challenges, such as the problem of identifying and addressing issues of jurisdiction for legal access to data and the lack of adequate channels for international communication and cooperation during cyber forensic investigations.

The first option, recovering evidence from seized local devices known to have interacted with the cloud, has several advantages. The devices can easily be accessed, and the cost of forensic analysis is relatively low. An exhaustive review of the client devices without accessing cloud servers can show some evidentiary artifacts useful in an investigation.

A substantial amount of research on client forensics has focused on the identification and analysis of the primary sources for historical evidentiary artifacts, resulting in large amount of data, from several sources, which investigators are not able to connect (Marturana, Me, & Tacconi, 2012; Quick & Choo, 2013; Quick, Martini, & Choo, 2014; Malik, Shashidhar, & Chen, 2015; Long, & Qing, 2015; Epifani, 2013).

There is lack of research that further processes identified historical artifacts to help the investigator determine the relationships among the created artifacts for more effective investigations.

The paper describes research work towards developing a Neo4j graph-based database tool which allows for prompt location of evidentiary artifacts with the goal to enhance client forensic analysis. The question this research attempts to answer is: *How can we leverage existing client forensics research and findings to build a tool the investigators can use to locate evidentiary artifacts promptly, given that one or multiple devices have been used to access a cloud storage service?*

The research scenario is as follows: *An organization suspects that documents containing designs for a new product have leaked to a competitor. The suspicion is that an insider might have used a cloud storage service to leak the material. The suspect devices, which include a Windows PC, MacBook, Android, and Windows phone, become target devices and are seized. The devices have some apps installed. As*

*an internal investigator, you want to quickly locate the evidentiary artifacts to prove cloud storage usage and be able to attribute actions to the suspect. There is a litigation hold in place.*

Identifying typical crime-related fingerprints is hard and the work proposed here contributes to speeding the process.

The outline for the paper is as follows. Section 2 presents the process of building the initial dataset from existing client forensics research. Section 3 provides a schema model for the Neo4j graph database and the rationale for choosing graph databases. Section 4 covers conclusions and future work.

## 2. BUILDING THE DATASET FOR THE DATABASE

The primary goal of this research is to augment existing research by developing a tool to timely locate evidentiary artifacts, given that one or multiple devices have been used to access a cloud storage service.

The architecture of the tool includes a Neo4j graph database, which contains for each identified cloud storage service and for each platform used to access the service, the likely types, and location of artifacts that constitute evidence of usage. The development of the tool involves:

- Identification the dataset to initially populate the database.
- Designing the data model based on the artifacts in a)
- Populating the database and implementing the interface.

The ultimate goal is to make this an Open Source project where researchers can contribute similar data, as they carry out investigations.

### 2.1 Building the Initial dataset

The building of the initial database to populate the database involves:

- •Assessing existing research on client forensics and providing for each commonly used storage service a collection of artifacts created by user activity, and platform used during the cloud storage interaction.
- • Based on a) for each storage service, identify a standard set of data artifacts, location, user activity, and platform used, as determined by several researchers to prove usage.

The research targets artifacts generated by the following user activities; (1) installation of a cloud service on a device used to access the cloud service; (2) uploading, downloading, moving, copying, and accessing of user data files; (3) uninstalling of the service; and (4) use of anti-forensic techniques (erasing the apps, data files, uninstalling the app).

The research focuses on commonly used cloud storage services such as Google Drive, Dropbox, and OneDrive through Android, Windows, iPhone, and Windows PC. A cloud storage service can be accessed either through the installed client or a browser (Internet Explorer, Mozilla Firefox, and Chrome).

Some of the notable research assessed involved accessing:

- Google Drive and Dropbox from a Windows 7 PC and an iPhone 3G (Quick & Choo, 2013; Quick & Choo, 2014) .
- Amazon S3, Dropbox, Evernote, and Google docs from Windows XP/Vista/7, a Mac PC, and an iPhone 3G (Chung, Park, Lee, & Kang, 2012).

- Dropbox, Google Drive, and SkyDrive from a Windows 7 PC and iPhone 3G (Epifani, 2013).
- Copy and ownCloud from Windows 8.1 PC (Malik, Shashidhar, & Chen, 2015).
- Google Documents, Flickr, PicasaWeb, Dropbox from a Windows 7 PC (Marturana, Me, & Tacconi, 2012).
- 360 and Baidu from a Windows 7 PC (Long, & Qing, 2015).

Some relevant conclusions from the assessment are as follows:

- a. There is a significant amount of collected artifacts, from various sources making it hard for law enforcement to figure relationships among existing data artifacts.
- b. Accessing a cloud storage service through a Web browser or client software creates a substantial number of artifacts that can be used to prove usage of the service. Examples of artifacts include the cloud storage service used, installation location, installed version, usernames, and passwords. These artifacts play an essential role in an investigation as they may lead an investigator to the possible position of other artifacts promptly.
- c. The identified artifacts depend on the browser used to access the storage. Ephani (2013) experimented with Mozilla Firefox and Internet Explorer. Chung used Internet Explorer, Quick and Choo used Mozilla Firefox, Google Chrome, Safari and Internet Explorer with Dropbox and Google Drive for access. Quick (Quick & Choo, 2013) findings noted that use of Mozilla and Google Chrome revealed a Google Drive account username through browser analysis. Use of Apple Safari did not show a username.

- d. As explained in the previous section, different sources of evidentiary artifacts were identified, depending on access mechanism. When a PC is used to access a service, the three principal sources of are the hard drive, the RAM and the eavesdropped network traffic between the device and the cloud network. When an iPhone 3G was used to access the service and a logical extract taken the specific locations for artifacts were database files, XML files, and plist files.
- e. The processes used by the researchers to identify artifacts are static and dynamic. The static approach, used by Quick (Quick & Choo, 2014; Quick & Choo, 2013) assumes that the investigator has a forensic image and can use forensic tools and prior acquired knowledge and skills to locate artifacts. The dynamic approach, used by Ephani (2013) and Malik (2015) use software tools, such as Disk Pulse and RegShot to locate the artifacts, while the experiment activity is underway, and the PC being used to access the service is on. Regardless of the approach used, similar data artifacts were identified.

Understanding how devices, information systems, and software interact and how they can be compromised, along with the types of evidential artifacts that may still be resident on those devices, has immediate and imminent impacts on both security and intelligence efforts both today and in the future (Muchmore & Duffy, 2017).

## 2.2 Designing the Graph database model

After identifying the initial dataset for the data, the next step is to provide the structure for the data, which is the data model. A model describes the domain as a connected graph and relationships.

A graph is composed of a node and a relationship. A node is an entity. A relationship represents a connection between nodes. The process of modeling the data include:

1. **Identifying the nodes, relationships, properties and labels from the problem domain.** A node is an entity with a unique conceptual identity which can have a relationship, a label, and properties. A relationship can have properties as well. A label is a graph construct used to group nodes into sets and has a name.

For the client forensics graph database, there are six nodes, each with the following labels:

- a. Service: The cloud storage service; the set includes Google Drive, OneDrive, and Dropbox.
- b. Artifact: All objects of digital archaeological interest (Forensic Wiki, n.d); examples include created registry keys.
- c. Platform: The type of operating system and the browser on the device used to access the storage service; can either be mobile (Android, Windows Mobile, or Windows PC). Browsers include Chrome, Internet Explorer, Safari, or Firefox. Created artifacts depend on the browser used to access the service.
- d. Activity: Activities that generated the artifact, the range of activities are outlined in the previous subsection.
- e. Source: The different sources of artifacts including the browser, client, and RAM, or network capture file, and mainly depends on the platform. Also, possible pathname for the source is presented.

2. **Identifying the interactions between the entities or nodes.** The following relationships were identified:
  - a. An artifact belongs to a service
  - b. An artifact is generated through an activity
  - c. An artifact has a source
  - d. A service is installed on a platform
  - e. A platform determines location of the artifacts
3. Draw the graph data model. Fig 1 shows the sketch of the data model representing the nodes and relationships. The sketch was drawn using apcjones Arrow Tool (Arrow Tool, n.d)
4. Figure 2 shows sample data to be loaded into the database.

### 2.3 Why choose a graph database?

There are several reasons to choose a graphical database. The artifacts identified from different sources are crucial for client forensic investigations and represent a massive data. The data artifacts are continually changing as operating system versions change. Graph databases provide the best means for modeling and make it easy to evolve according to changes in operating system versions. The original data remains intact, while new nodes and relationships are added.

Once there are significant data, which is continually changing, traditional databases are inadequate regarding response time and are do not show good performance when applied to large datasets. Storing data relationship directly as a graph, made up of nodes or vertices, reduces complexity and eliminates the extra work involved in transforming the data from the model to storage.

Neo4j has been known to improve application performance. There are several companies which are currently using Neo4j database systems; examples include eBay which is speeding e-commerce delivery routing using Neo4j. eBay's same-day delivery grew exponentially, and its service platform needed a revamp to support the explosive growth in data and new features (Neo4j Graph Database Platform, 2018). The MySQL joins created a code base too slow and complicated to maintain the queries used to select the best carrier were taking too long. eBay picked the Neo4j for its flexibility, speed, and ease of use.

Neo4j databases have found their way into fraud detection, as well. The traditionally used fraud detection measure focus on data points such as specific accounts, individuals, and devices or IP addresses. Today, fraudsters are forming fraud a ring of stolen and synthetic identifies. To uncover the fraud rings, there is need to look more at the connections that link identifies. Neo4j has been known to detect patterns that far outstrip the power of a relational database (Neo4j Graph Database Platform, 2018).

#### **2.4 How the investigator will benefit from using this tool?**

After the modeling of the database, an interface will be implemented using Python. The developed tool facilitates investigators' understanding of complex relationships among various data artifacts, over and above traditional relational databases. Using the designed interface and the provided scenarios, the investigators can use drop-down menus to look up information such as:

What installation activities were carried out on each device?

What user activities associated with the cloud service were performed on each device?

What is the timeline for activities performed and artifacts created artifacts on all the devices?

### **3. CONCLUSION**

The main conclusion from the presented work is that some evidentiary artifacts, useful in an investigation are obtainable through an exhaustive analysis of the client devices, without accessing the cloud server. Building a tool for timely identification of these artifacts enhances the investigation process.

The growing popularity of cloud storage services means that this media will be used for cybercrime, resulting in more investigation cases. One challenge is maintaining a chain of custody in the cloud; there is a need for more research in this area.

Future work should include accessing popularly used cloud storage services from commonly used mobile platforms; Android, Windows, and the latest iPhone. A series of experiments that involve installing, accessing, uploading and downloading some documents; uninstall the client software, and then using anti-forensics techniques (deletion, uninstalling and clearing the browser history) to hide usage is being performed for the popularly used cloud services, accessed from Android. For each mobile device. We have started work on rooting an Android device, to collect the primary system folders.

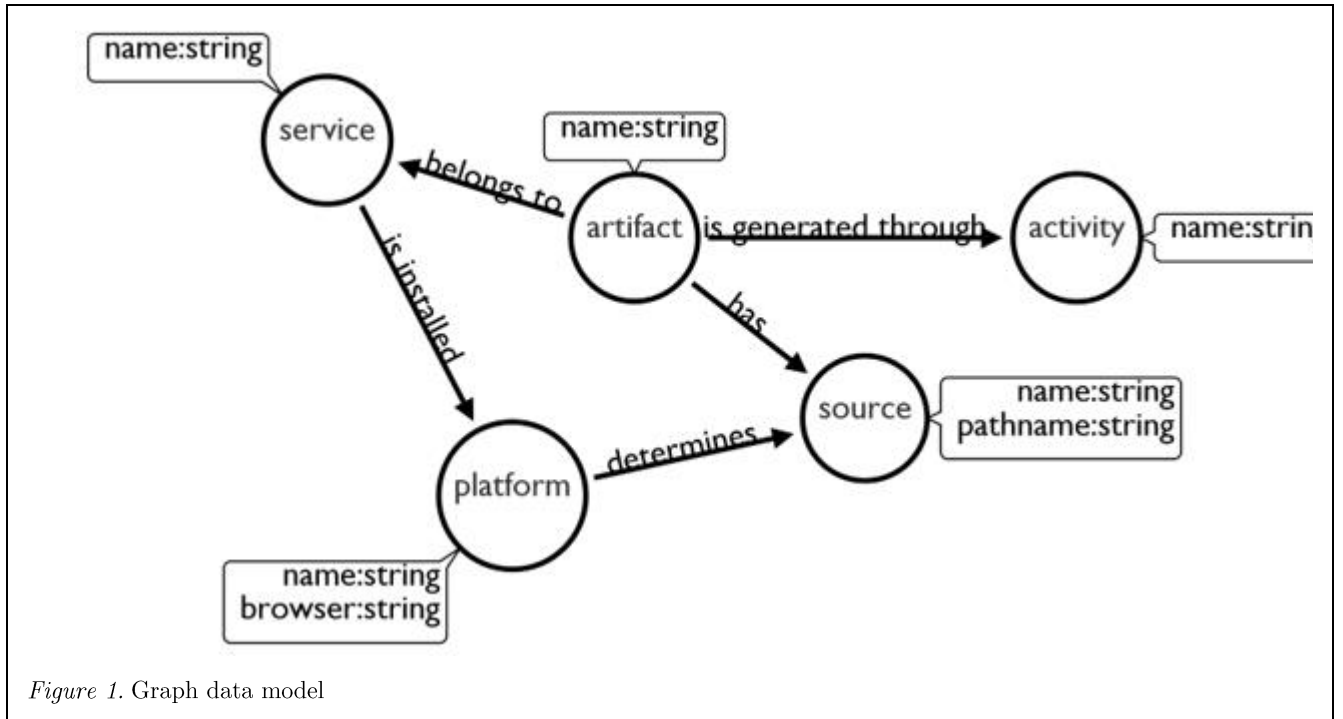


Figure 1. Graph data model

Service	Platform	Activity	Artifact	Source
Google Drive	Windows7	installation	path for the syncing folder	client - user profile -sync_confid.db
Google Drive	Windows7	installation	user email for Google Drive account	client - user profile -sync_confid.db
Google Drive	Windows7	installation	downloaded software	browser - C:\Users\<User_Name>\AppData\Local\Google\Chrome\UserData\Default\downloads

Figure 2. Typical dataset to be populated into the database



## REFERENCES

- The Arrow Tool: Retrieved on January 31, available here: <http://www.apcjones.com/arrows/>
- Cloud Storage Report 2017 – Dropbox Loses Market Share But is Still the Biggest Provider on Mobile: Retrieved on January 31, 2017 from: <https://blog.cloudrail.com/cloud-storage-report-2017/>
- Chung, H., Park, J., Lee, S., & Kang, C. (2012). Digital forensic investigation of cloud storage services. *Digital Investigation*, 9(2), 81-95.
- Dardick, Natalie; ADFSL, Baggili, Ibrahim; Zayed University, Carthy, Joe; University College Dublin, & Tahar; University College Dublin (Eds.). (2011). Survey on Cloud Forensics and Critical Criteria for Cloud Forensic Capability: A Preliminary Analysis. ADFSL.
- Epifani, M. (2013). Cloud Storage Forensics. Paper presented at SANS European Digital Forensics Summit, Prague.
- Faheem, Kechadi, Khan. (2014), An Overview of Cloud base Applications Forensics Tools for Mobile Devices, *International Journal of Applied Information Systems(IJAIS)*, Foundations of Computer Science, FCS, NY, USA, Volume 7-10
- Forensic Wiki: Retrieved on January 31, 2018, and available here: [http://forensicswiki.org/wiki/Computer\\_forensics](http://forensicswiki.org/wiki/Computer_forensics)
- Gartner, Cloud Adoption and Risk report. Retrieved on January 31, from here: <https://www.skyhighnetworks.com/cloud-report/>
- Grispos, G., Storer, T., & Glisson, W. B. (2012). Calm Before the Storm. *International Journal of Digital Crime and Forensics*, 4(2), 28-48.
- Zatyko & Bay, J (2011). The Digital Forensic Cyber Exchange Principle, *Forensic Magazine* (12)
- Long, C., & Qing, Z. (2015). Forensic Analysis to China's Cloud Storage Services. *International Journal of Machine Learning and Computing*, 5(6), 467-470.
- Malik, R., Shashidhar, N., & Chen, L. (2015). Analysis of Evidence in Cloud Storage Client Applications on the Windows Platform. Paper presented at Int'l Conf. Security and Management.
- Millis, E. (2012) "Cybercrime moves to the cloud" Retrieved on January 31, from: <http://www.cnet.com/news/cybercrime-moves-to-the-cloud/>
- Marturana, F., Me, G., & Tacconi, S. (2012). A Case Study on Digital Forensics in the Cloud. 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery.
- Muchmore, M, Duffy, J. The Best Cloud Storage and File Sharing Services of 2017, PC Magazine, March, 31, 2017: Accessed on June 29<sup>th</sup> from: <http://www.pcmag.com/roundup/306323/the-best-cloud-storage-providers-and-file-syncing-services>
- Narayan, Kaushik. (2016) 5 devious Instances of Insider Threat in the Cloud; Retrieved on January 31, from: <https://www.skyhighnetworks.com/clou>

- [d-security-blog/5-devious-instances-insider-threat-cloud/](#)
- NIST. (2014). NIST Cloud Computing Forensic Science Challenges (NISTIR 8006).
- Quick, D., & Choo, K. R. (2013). Dropbox analysis: Data remnants on user machines. *Digital Investigation*, 10(1), 3-18.
- Quick, D., & Choo, K. R. (2014). Google Drive: Forensic analysis of data remnants. *Journal of Network and Computer Applications*, 40, 179-193.
- Quick, D., Martini, B., & Choo, K. R. (2014). Forensic Collection of Cloud Storage Data. *Cloud Storage Forensics*, 153-174.
- Ruan, K., Carthy, J., Kechadi, T., & Baggili, I. (2013). Cloud forensics definitions and critical criteria for cloud forensic capability: An overview of survey results. *Digital Investigation*, 10(1), 34-43. Stamford. (n.d.). Gartner Says That Consumers Will Store More Than a Third of Their Digital Content in the Cloud by 2016. Retrieved from <http://www.gartner.com/newsroom/id/2060215>
- Ruan, Keyun; Baggili, Ibrahim; Carthy, Joe; and Kechadi, Tahar, "Survey on Cloud Forensics and Critical Criteria for Cloud Forensic Capability: A Preliminary Analysis" (2011). Annual ADFSL Conference on Digital Forensics, Security and Law. 2. <https://commons.erau.edu/adfsl/2011/friday/2>
- Six security risks of enterprises using cloud storage and file sharing apps; Retrieved on January 31, 2017 from <https://digitalguardian.com/blog/6-security-risks-enterprises-using-cloud-storage-and-file-sharing-apps>
- Taylor, M., Haggerty, J., Gresty, D., & Hegarty, R. (2010). Digital evidence in cloud computing systems. *Computer Law & Security Review*, 26(3), 304-308.
- The dirty dozen: 12 cloud security threats; InfoWorld. Retrieved on January 31, from here: <http://www.infoworld.com/article/3041078/security/the-dirty-dozen-12-cloud-security-threats.html>
- The state of cloud computing: 10 things you need to know: Retrieved on January 31, from : <http://www.techrepublic.com/article/the-state-of-cloud-computing-10-things-you-need-to-know/>
- Top 10 Security Concerns for Cloud-Based Services. Retrieved on January 31, 2017 from: <https://www.incapsula.com/blog/top-10-cloud-security-concerns.html>
- <https://neo4j.com/developer/guide-data-modeling/>. accessed: 2017.16.04
- [Sasaki, Bryce Merki: Graph Databases for Beginners: Why Graphs are the Future: retrieved April 11<sup>th</sup>, 2018](#)
- Neo4j Graph Database Platform. (2018). Retail & Neo4j: Ecommerce Delivery Service Routing. [online] Available at: <https://neo4j.com/blog/retail-neo4j-ecommerce-delivery-service-routing/> [Accessed 15 Apr. 2018].
- Neo4j Graph Database Platform. (2018). Retail & Neo4j: Ecommerce Delivery Service Routing. [online] Available at: <https://neo4j.com/blog/retail-neo4j-ecommerce-delivery-service-routing/> [Accessed 15 Apr. 2018].

