NASA

Do Data Almost Always Eventually Leak?

Norita Ahmad, American University of Sharjah

Shannon's information theory states that there is a maximum rate for transmitting data reliably, implying that data almost always inevitably leak. Although there has been advancement in data protection technologies, a lack of understanding about data leaks and human error persists.

Claude Shannon, known as the father of modern digital communications and information theory, proposed the idea that once information became digital, it could be transmitted without error. Shannon's information theory, published in 1948,[1] showed that any given communications channel has a maximum capacity for reliably transmitting information. In other words, if one were to send information at a rate greater than the threshold, one will always lose part of the message. This result suggests the idea that all information leaks.

INFORMATION AND DATA

Now let us take a step back. What is information? From an information systems point of view, information refers to data that have been organized so that they have meaning and value to the recipient. Data, on the other hand, refer to a stream of raw facts that does not convey any specific meaning.[2] Alternatively, we can define information in the following way: X is an A, where X represents a piece of data, and A is something that explains that piece of data.[3]

Although we can process data into information, the reverse does not apply. Information is not reducible to something as self-explanatory as data.[3] According to Shannon, units of information can be defined as a *bit* or *binary digit*, which is the smallest possible chunk that cannot be divided any further.[1] Strings of bits are used to encode messages. Therefore, when someone says that information has leaked, what exactly does that mean? Does it mean that someone has successfully transferred the bits representing the information to an external

Digital Object Identifier 10.1109/MC.2020.3041880
Date of current version: 11 February 2021

0018-9162/21©2021IEEE

EDITORS
NORITA AHMAD American University of Sharjah, nahmad@aus.edu
PREETI CHAUHAN VP Technical Activities, IEEE Reliability Society, preeti.chauhan@ieee.org


unauthorized user? Does it also mean that the bits contain all there is to the information? Drawing on the previous definitions of data and information, if someone were to obtain the data, it does not necessarily mean that she or he has the information. It is possible to bury the signal used for data transmission in the noise so that only those who are the intended receivers of the data know of the presence of the signal and how to extract it from the noise.

Interestingly, one would then ask, what does it mean for data to leak? A data leak refers to data that were made available to unauthorized people intentionally or unintentionally.[4] For example, an insider could make an unauthorized disclosure of how his or her organization transmits or stores data using a particular steganographic technique. In the context of privacy, a data leak means the *further dissemination of data beyond what is permitted by privacy policy*.[4] Given the fact that society has become increasingly reliant on digital data, cloud computing, e-commerce, e-government services, and workforce mobility, individuals need to understand the risks posed by data leaks. The risk depends on the specifics of the scenario. Consider, for example, the usage, storage, transmission, and processing of contact-tracing data during the COVID-19 pandemic. Data leakage from the contact-tracing system may not be frequent or severe if the information used for mapping those data to specific persons is in a separate system and accessible only to authorized users. However, it has been shown that data inferencing and aggregation techniques can be used to construct the mapping between contact-tracing data and the individuals to which those data refer.[5]

Now let us take a look at the millions of users who share data via social media platforms such as Facebook and Twitter or the ever-growing number of records collected and analyzed by cloud service and applications providers such as Amazon and Google. The reality is that these companies might collect far more personal data about their users than most people realize. For example, Google records every search that is performed on the Google search engine, every video watched on YouTube, every place visited, and every route taken using Google Map. Google basically knows everything about everyone. What if these data were accidentally exposed, like the series of data leaks from Amazon S3 buckets[6] that exposed users' sensitive data to the Internet? These data leaks were caused by simple errors, but the impact was huge on both the company and users.

An important point here is that a cyberattack is not a prerequisite for data leakage. A data leak can stem from poor data security practices or accidental actions by an individual or a group of people.[8] When data are processed, they usually flow through a chain of services or points[1] (such as a human being, computers, servers, or cloud services) that can cause a data leak.[4] In general, poor application security and measures in any part of the chain can cause a leak.

DATA SECURITY

With the rise of Internet use, online shopping, and other online activities, personal data are becoming highly valued. In a recent study done by Varonis, a data security company based in New York, it was reported that many companies did not follow a strict protection of customers' personal data, especially from anyone inside the company. On average, out of 785 organizations from more than 30 industries worldwide, it was found that 53% of the companies had more than 1,000 sensitive files accessible to all employees in the organization, and only 5% of folders in all of these companies were properly protected.[7] Even though this phenomenon is not new with digital systems, the risks are just much more widespread now. That is why the security and privacy communities have invested so much effort in devising cryptographic means for securely storing and transmitting data.

Digitization is fundamentally affecting every individual in the world. One might say that only e-commerce companies should worry about this because they are in the business of data and must therefore comply with the local laws or policies of where they are headquartered or do business. However, even if you are not transacting any business online and are only dealing with physical goods or providing a service such as appliance repair, you still generate a lot of data.

What if you were to outsource data processing to a third party? Even if you were to have top-notch security tools, prevention, and protection in place, the companies that are processing your data may not, which could still compromise your data security. Worse still, even if you were to use cloud services or servers on Azure or Amazon, often perceived as highly secure, data leaks can occur. Although there are general guidelines, laws, and regulations in the United States, such as the Health Insurance Portability and Accountability Act and the Gramm–Leach–Bliley, Sarbanes–Oxley, and Family Educational Rights and Privacy Acts, which must be followed by all organizations, the way data are handled differs from business to business or even from industry to industry.

Ultimately, it is up to individual organizations and their employees to

follow standards in their daily operations. The danger is when businesses do not have enough awareness or visibility of how their data are actually being stored or handled. That is why most data leaks are unintentional operational problems rather than strictly traditional cybersecurity problems.[8] In the case of Amazon, unfortunately, it was an intentional act by its employees who were immediately fired for leaking customer data from Amazon S3 buckets to an unaffiliated third party in violation of company policies.[9]

Although there are best practices to follow to reduce the likelihood of a leak, nothing is foolproof. As evident in the Amazon case, it is extremely difficult to guard against either human error or insider malicious intent regardless of all the security precautions and data-leak-prevention practices followed by an individual or organization. The problem is that the prevention is only as good as the humans and policies that govern it. For example, if the policy for data-leak prevention is outdated and has not been updated to reflect the current workforce, then many would have issues accessing any data.[8] Similarly, if the policies do not allow for cross-departmental sharing, then any data sharing among employees from different departments would be considered data leakage. In this scenario, the employees may not understand the reason why they could not share data; therefore, there needs to be effective security and awareness training.[10]

PROTECTION EFFORTS

To prevent more leaks of personal data from happening in the future, the European Union introduced the General Data Protection Regulation (GDPR) on 25 May 2018.[11] The main purpose of GDPR is to give users full control over their own personal data collected by companies. In addition, GDPR also requires companies to report any potential breach of data to the country's data protection authority within 72 h of the incident; failure to do so would result in GDPR fines.[11] An important lesson for everyone to learn here is that no matter how much the company invests in security tools and how many cybersecurity precautions a company takes, there is no guarantee that those data can be protected forever. GDPR was initiated for this very reason, that is, to help minimize the exposure and exploitation of personal data that are stored online by requiring companies not only to allow users to control their personal data but to establish time limits for data storage.[11]

> A known problem with cloud storage is that users do not know how their data are being protected from compromise.

Today, many companies such as Google, Facebook, and Microsoft provide options for users to control their personal data collected by the company. For example, people can restrict browsers from sending location details to sites visited, remove super cookies and other cookies, and notify Google to delete private data and limit how long the company holds onto those data (the default limit is 18 months). Recently, Google announced that users can now opt out entirely from all of the smart features offered by the company, such as Smart Reply.[12] Does it now mean that users are in better control of their personal data and the data collected about them? Does it now mean that it will be less likely for private data to leak? Now ponder this: we might have taken all of the necessary precautions and actions to keep a certain level of privacy, but what about the mutual friends and family members on our social network sites? These people will also determine our privacy. So, is it fair to say that data almost always will eventually leak?

In the case of the revelation of information, a leak could simply be the description of the characteristic of information itself, highly dependent on whether the contents are rightfully contained.[13] It is known that anything in a state of containment is prone to leak, including the flows of data and information.[13] A well-known example of this is the 2010 release of U.S. diplomatic cables by WikiLeaks, where highly classified information was disclosed to the general public. Heather Brooke stated, "Leaks are not the problem, they are the symptom," and added that the Wikileaks incident shows a disconnect between what people want to know, need to know, and actually do know.[14] She further argued that this disconnect is not always a sign of deficiency but rather the result of a system designed to contain.

Given the relational nature of leaks—where they come out of something, someone, or somewhere into something, someone, or somewhere else[13]—we could then argue that the greater the secrecy, the more likely it is to leak. This could be one of the reasons why people keep leaking government secrets. Although classifying information is a key part of any government, some have questioned whether too much government information gets labeled unnecessarily as such, and, as it turns out, the Wikileaks scandals led many to believe so.[15] In 1971, while referring to the leaking of the Pentagon Papers case, Supreme Court Justice Potter Stewart was famously quoted for saying, "When everything is classified then nothing is classified."[16]

SECURITY IN THE DIGITAL AGE

In the world of the digital age, it becomes easier for anything to be shared. Accordingly, one quick solution to the disclosure of confidential government information is to treat information the

way it deserves to be treated, instead of keeping it hidden under wraps.[15] We have already experienced the massive impact of digitization, either positively or negatively. Media technologies have evolved significantly over the period between the leaking of the Pentagon Papers in 1971 and the occurrence of U.S. diplomatic cable leaks in 2010. Information that was once considered private is now aggregated and can potentially become permanently public. The ability to reproduce, share, store, move, and disperse information becomes increasingly easy and more efficient, resulting in the exponential growth of information itself, thus affecting the amount of information that can then be leaked. In addition, compared to the leaks of the previous era, such as the Pentagon Papers, today, the public has access to and can interact with the content of the leak.[16] Technology has enabled people to read, see, and hear everything that is relevant to them. It is therefore no surprise that the reaction to information leaks is different today compared to before.

As the world moves in the direction of openness, people need to understand the difference between secret and nonsecret data. Disclosing data can be a double-edged sword: it may provide an informational benefit while, at the same time, enabling the leakage of private information and even damaging national or global security. Although the problem of privacy and information leakage has been around for a long time, we are still struggling to keep it under control. Digitization is breaking down the traditional barriers of exclusion and replacing it with an ethos of collaboration and transparency.

Going back to Shannon's information theory, he showed that one of the advantages of a digital system was the fact that we could choose how we represent information with bits so that information can basically exist in a variety of situations. Not only that, Shannon's basic bits made it possible for us to create a simpler and cheaper representation of information today.[1] The implication though, with bits, is that people can also manipulate, copy, alter, and share information endlessly. As such, the role and interest in information theory will continue to grow, as it provides fundamental insights into a better understanding of how much information leaks and to what extent it can be reduced and tolerated.[17] We have also seen technical solutions proposed in *Computer*, such as a means for empowering users to take control of their data regardless of where they are stored or over what communication infrastructure they are transmitted.[18]

The most important takeaways from this article are

- *The danger of insider threats*: We should never take internal threats lightly, as many major data leaks have been linked to insider threats.
- *Do not always trust your personal network*: We should remain skeptical of people around us, including family and friends, as their lack of security awareness could cost us our privacy.
- *The vulnerability of cloud storage*: As discussed in the Amazon S3 buckets case, in addition to the risk of cyberattacks, there is an increased risk of insider threats. A known problem with cloud storage is that users do not know how their data are being protected from compromise.
- *That compliance alone is not enough*: Although most companies have sought compliance, many failed to understand that compliance helps achieve only the bare minimum. Companies should focus more on the pivotal roles of employees' involvement in protecting information as well as more comprehensive and up-to-date policies.
- *That data almost always inevitably leak because of human error*: It is important for us to understand that even with the advancement of technology and the strength of cryptographic algorithms, the weakest links are still humans.

ACKNOWLEDGMENTS

I thank Prof. Bret Michael and Rick Kuhn for their edits on earlier versions of this article.

REFERENCES

1. C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.
2. K. C. Laudon and J. P. Laudon, *Management Information Systems: Managing the Digital Firm*. London: Pearson Education Ltd., 2018.
3. J. Barwise and J. Seligman, *Information Flow: The Logic of Distributed Systems*. Cambridge, U.K.: Cambridge Univ. Press, 1997.
4. C. Colon. "What's the difference between a data leak and data breach?" SandStormIT. 2019. https://sandstormit.com/whats-the-difference-between-a-data-leak-and-data-breach/ (accessed Nov. 12, 2020).
5. N. Ahmad and P. Chauhan, "State of data privacy during COVID-19," *IEEE Ann. Hist. Comput.*, vol. 53, no. 10, pp. 119–122, 2020. doi: 10.1109/MC.2020.3010549.
6. A. Scroxton. "Exposed AWS buckets again implicated in multiple data leaks." ComputerWeekly. Jan. 20, 2020. https://www.computerweekly.com/news/252476870/Exposed-AWS-buckets-again-implicated-in-multiple-data-leaks (accessed Nov. 14, 2020).
7. R. Sobers. "2019 data risk report stats and tips you won't want to miss Varonis." Cyber Security News. June 17, 2020. https://www.varonis.com/blog/data-risk-report-highlights-2019/ (accessed Nov. 22, 2020).
8. R. Layland, "Data leak prevention: Coming soon to a business near you,"

Bus. Commun. Rev., vol. 37, no. 5, p. 44, 2007.
9. K. Afifi-Sabet, "Amazon sacks employee over data breach." ITPro. Oct. 27, 2020. https://www.itpro.co.uk/security/357555/amazon-data-breach-sacks-employee (accessed Nov. 14, 2020).
10. J. Haney and W. Lutters, "Security awareness training for the workforce: Moving beyond 'check-the-box' compliance," *Computer*, vol. 53, no. 10, pp. 91–95, 2020. doi: 10.1109/MC.2020.3001959.
11. "General Data Protection Regulation (GDPR)." Intersoft Consulting. https://gdpr-info.eu (accessed Nov. 22, 2020).
12. J. Porter. "Google will soon let you opt out of Gmail's data hungry smart features entirely." The Verge. Nov. 16, 2020. https://www.theverge.com/2020/11/16/21569599/google-gmail-meet-chat-personal-data-toggle-privacy (accessed Nov. 17, 2020).
13. D. Clark, S. Hunt, and P. Malacaria, "Quantitative analysis of the leakage of confidential data," *Electron. Notes Theoretical Comput. Sci.*, vol. 59, no. 3, pp. 238–251, 2002. doi: 10.1016/S1571-0661(04)00290-7.
14. H. Brooke, (2010). WikiLeaks: The revolution has begun – and it will be digitized." The Guardian. www.guardian.co.uk/commentisfree/2010/nov/29/the-revolution-will-be-digitised?INTCMP=SRCH (accessed Nov. 12, 2020).
15. M. Giglio, "The U.S. government keeps too many secrets." The Atlantic. Oct. 3, 2019. https://www.theatlantic.com/politics/archive/2019/10/us-government-has-secrecy-problem/599380/ (accessed Nov. 12, 2020).
16. N. Colvin. "The logic of leaks, reconsidered Limn." 2017. http://limn.it/the-logic-of-leaks-reconsidered
17. G. Smith, "Quantifying information flow using min-entropy," in *Proc. 2011 8th Int. Conf. Quan. Eval. SysTems*, pp. 159–167. doi: 10.1109/QEST.2011.31.
18. J. Michael, "Empowering users through secure on-demand data provisioning," *Computer*, vol. 46, no. 6, pp. 84–85, 2013. doi: 10.1109/MC.2013.203.

NORITA AHMAD is an associate professor of management information systems at American University of Sharjah, Sharjah, United Arab Emirates. Contact her at nahmad@aus.edu.