Journal of the Midwest Association for Information Systems

Table of Contents

Volume 2024, Issue 1, January 2024 www.JMWAIS.org

Journal of the Midwest Association for Information Systems

Volume 2024 | Issue 2 Article 1

Date: 07-02-2024

Societal Impact and Accreditation: Implications for Information Systems Research and Practice

Barbara D. Klein
University of Michigan-Dearborn, bdklein@umich.edu

Rassule Hadidi
Metro State University, Rassule.Hadidi@metrostate.edu

Abstract

The *Journal of the Midwest Association for Information Systems* has, since its inception, played a role in publishing manuscripts focused on positive societal impact in the Midwest and beyond. This paper discusses current AACSB, ACBSP, and ABET accreditation standards focused on societal impact and offers examples of manuscripts published in the journal categorized by the United Nations Sustainable Development Goals which is suggested as a taxonomy for societal impact by AACSB accreditation standards. Suggestions for future manuscripts are then offered to interested information systems faculty who are seeking an outlet for manuscripts reporting their curricular, scholarly, and other activities with societal impact. Potential focus areas suggested for future work include the elimination of the digital divide, measuring and reporting social mobility index, the use of information systems to support the development of resilient agriculture, the use of information systems to monitor and improve water quality, high impact teaching practices and information systems curriculum innovations, gender equality in information systems education and practice, eliminating or at least reducing the learning equity gap, and remote and hybrid work arrangements.

Keywords: societal impact; social mobility index; information systems; accreditation standards; AACSB; ACBSP; ABET; gender equality; economic growth; infrastructure; high-speed Internet access; digital divide; resilient agriculture; water quality; pedagogy; remote work; United Nations Sustainable Development Goals

DOI:10.17705/3jmwa.000088

1. Introduction

Current accreditation standards addressing information systems programs in colleges of business, management, and engineering focus increased attention on the societal impact of research, teaching, and service initiatives. This paper will discuss societal impact with respect to AACSB, ACBSP, and ABET accreditation standards, highlight key areas on which information systems faculty may wish to focus their work to address societal impact in accreditation standards, and outline ways in which the *Journal of the Midwest Association for Information Systems* can support societal impact initiatives and serve as an outlet for faculty interested in publishing reports of their work related to societal impacts.

2. Accreditation Standards and Social Impact

Although scholars and practitioners working in the information systems field have traditionally been concerned with the individual, organizational, and societal impacts of their work (e.g., Chen et al, 2008; Clemons et al, 2016; De Leoz et al., 2018; Dwivedi et al., 2023; Malhotra et al., 2013; Melville, 2010; Sabherwal & Grover, 2024), current accreditation standards governing information systems programs in college of business, management, and engineering make these perspectives and goals explicit.

2.1 AACSB Accreditation Standards and Societal Impact

Many information systems programs located in colleges of business and management are accredited by AACSB or ACBSP. The "2020 Guiding Principles and Standards for Business Accreditation" place an explicit and increased emphasis on societal impact (AACSB, 2023). The standards require business and management schools to articulate how they have a positive societal impact across several standards that must be addressed in order for a college of business and management to attain or retain accreditation (AACSB, 2023). Specifically, standard 9 states that: "The school demonstrates positive societal impact through internal and external initiatives and/or activities, consistent with the school's mission, strategies, and expected outcomes." Standard 8 also states that: "portfolio of intellectual contributions contains exemplars of research and publications that have a positive societal impact that is consistent with the school's mission and strategic plan," (AACSB International 2020 Standards).

Additional guidance for setting and measuring goals related to societal impact is provided in a supplemental document titled "AACSB and Societal Impact: Aligning with the AACSB 2020 Business Accreditation Standards" (AACSB, February 2023).

Given the collective role business and management school faculty play in advancing and supporting positive societal impact, information systems faculty are understandably concerned with how their teaching, research, and service activities can advance accreditation standards related to societal impact. AACSB standards require curriculum and intellectual contributions that help the school achieve societal impact with a focus on positive societal impact stated in the school's strategic plan. The standards welcome societal impact that is local, regional, national, and international in scope, with activities that achieve social impact typically involving engagement with stakeholders relevant to the business and management school's mission (AACSB, February 2023).

AACSB standards allow, but do not require, activities with societal impact to be categorized and reported using the United Nations Sustainable Development Goals which are listed in Table 1 below (United Nations, 2024).

2.2 ACBSP Accreditation Standards and Societal Impact

Information systems programs located in colleges of business and management may be accredited by ACBSP which also has some focus on societal impact. ACBSP standards emphasize the consideration of impact on society as students learn to design solutions to technical problems. Specifically, standard one of the ACBSP discusses "Social and Community Responsibility" and "Impact on Society" (ACBSP, February 2024) which requires accredited institutions to "Describe the processes used by the business unit's leadership to identify and address the impact on society of its program offerings, services, and operations," (ACBSP, February 2024). The Criterion 5.3.C of standard five of ACBSP (faculty focus) focusses on scholarship using the Boyer Model. Under the scholarship of integration section, the standard further elaborates that "It is essential to integrate ideas and then apply them to the world in which we live," (ACBSP, February 2024).

	Goal
1	No Poverty
2	Zero Hunger
3	Good Health and Well-Being
4	Quality Education
5	Gender Equality
6	Clean Water and Sanitation
7	Affordable and Clean Energy
8	Decent Work and Economic Growth
9	Industry, Innovation, and Infrastructure
10	Reduced Inequalities
11	Sustainable Cities and Communities
12	Responsible Consumption and Production
13	Climate Action
14	Life Below Water
15	Life on Land
16	Peace, Justice and Strong Institutions
17	Partnerships for the Goals

Table 1. Sustainable Development Goals (United Nations, 2024)

2.3 ABET Accreditation Standards and Societal Impact

Information systems programs located in colleges of engineering may be accredited by ABET which also has a focus on societal impact. ABET standards emphasize the consideration of impact on society as students learn to design solutions to technical problems. Curriculum standards for computing programs accredited by ABET require coverage of "local and global impacts of computing solutions on individuals, organizations, and society" (ABET, 2024). While not identical to AACSB, or ACBSP standards, the suggestions in the following section should be helpful as information systems faculty working in colleges of engineering consider ways to publish manuscripts focused on societal impact in the *Journal of the Midwest Association for Information Systems*.

3. Societal Impact and the *Journal of the Midwest Association for Information Systems*

Given the critical role information systems play in shaping and supporting societies around the world, information systems faculty have the potential to support the attainment of initiatives focused on societal impact in the context of accreditation standards. The journal has published many papers and provided an outlet since its inception for papers reporting intellectual contributions and engagements with stakeholders that have societal impact. Example manuscripts published in the journal categorized by selected United Nations Sustainable Development Goals (United Nations, 2024) are presented in Table 2 below.

Goal	Topic	Manuscript
Zero Hunger	Agricultural Practice	Power, D., and Hadidi, R. (2019). Transforming agriculture: Exploring precision farming research needs. *Journal of the Midwest Association for Information Systems*. 2019(2), Article 1.
Good Health and Well-Being	Pandemic Preparedness	George, J. F., and Hadidi, R. (2022). COVID-19 and its impact on the Midwest United States. *Journal of the Midwest Association for Information Systems*. 2022(1), Article 1.
	Health Information Systems	Heavin, C. (2017). Health information systems – Opportunities and challenges in a global health ecosystem. *Journal of the Midwest Association for Information Systems*. 2017(2), Article 1
	Global Health Initiatives	Fernández, E. (2017). Innovation in healthcare: Harnessing new technologies, *Journal of the Midwest Association for Information Systems*. 2017(2), Article 8. Kenny, G., O' Connor, Y., Eze, E., and Heavin, C. (2017). Trends, findings, and opportunities: An archival review of health information systems research in Nigeria. *Journal of the Midwest Association for Information Systems*. 2017(2), Article 6.
	Healthcare Workers and Information Systems	Vroegindeweij, R., and Carvalho, A. (2019). Do healthcare workers need cognitive computing technologies? A qualitative study involving IBM Watson and Dutch

		professionals. *Journal of the Midwest Association for Information Systems*. 2019(1), Article 4.
Quality Education	High Impact Teaching Practices	Eierman, M. A., and Iversen, J. H. (2018). Comparing test-driven development and pair programming to improve the learning of programming languages. *Journal of the Midwest Association for Information Systems*. 2018(1), Article 3. Hadidi, R., and George, J. F. (2022). COVID-19 and examples of "best" teaching practices from the lens of different stakeholders. *Journal of the Midwest Association for Information Systems*. 2022(2), Article 1. Lebens, M. (2021). Using prototyping to teach the design thinking process in an asynchronous online course. *Journal of the Midwest Association for Information Systems*. 2021(2), Article 3. Luse, A., and Burkman, J. (2018). Safely using real-world data for teaching statistics: A comparison of student performance and perceived realism between dataset types. *Journal of the Midwest Association for Information Systems*. 2018(1), Article 2. Mitchell, A. J. D. (2018). Small business website development: Enhancing the student experience through community-based service learning. *Journal of the Midwest Association for Information Systems*. 2018(2), Article 4. Hadidi, R., and George, J.

		F. (2023). Potential uses of AI-based platforms in teaching and learning. *Journal of the Midwest Association for Information Systems*. 2023(2), Article 1.
Gender Equality	Gender and Use of Information Systems	Witt, C., Melton, J., and Miller, R. E. (2024). Gender, emotional intelligence, and the need for popularity: Exploring the causes of faux pas posting beyond the behavior of friends. *Journal of the Midwest Association for Information Systems*. 2024(1), Article 3. Kiely, G. L., Heavin, C., and Lynch, P. (2019). Building a shared understanding of female participation in IT through collaboration: A shared mental model approach. *Journal of the Midwest Association for Information Systems*. 2019(1), Article 3.
	Gender and the IT Workforce	Rowland, P., and Noteboom, C. B. (2018). Anchoring female millennial students in an IT career path: The CLASS anchor model. *Journal of the Midwest Association for Information Systems*. 2018(2), Article 3.
Decent Work and Economic Growth	Information Systems and Economic Development	Hadidi, R., Power, D., and George, J. F. (2016). Information technology is transforming the heartland: Making the case for Midwest United States. *Journal of the Midwest Association for Information Systems*. 2016(1), Article 1.
	Information Systems Career Preparation	Muraski, J. M. (2023). IS career day in a class: Raising college student awareness and interest in information systems.

		Journal of the Midwest Association for Information Systems. 2023(1), Article 3.
	Development of Degree Programs Targeting High Demand Career Fields	Strader, T. J., and Bryant, A. (2018). University opportunities, abilities and motivations to create data analytics programs. *Journal of the Midwest Association for Information Systems*. 2018(1), Article 4. Muraski, J.M., and Iversen, J. (2022). Growing computer science and information technology education in K-12: Industry demand and ecosystem support. *Journal of the Midwest Association for Information Systems*. 2022(2), Article 2. Muraski, J.M., Iversen, J., and Iversen, K.J. (2021). Building collaboration networks and alliances to solve the IT talent shortage: A revelatory case study. *Journal of the Midwest Association for Information Systems*. 2021(1), Article 3.
Industry, Innovation, and Infrastructure	Universal High-Speed Internet Access	Marett, K., and Xiao, S. (2022). Broadband Internet access as a localized resource for facilitating information security knowledge. *Journal of the Midwest Association for Information Systems*. 2022(1), Article 2.

Table 2. Examples of the *Journal of the Midwest Association for Information Systems* Manuscripts Focused on Selected Sustainable Development Goals (United Nations, 2024)

The *Journal of the Midwest Association for Information Systems* invites manuscripts across the spectrum of issues related to information systems and societal impact. We suggest that potential authors consult the United Nations Sustainable Development Goals taxonomy (United Nations, 2024) as well as other categories of societal impact that may be adopted by their institutions. We anticipate that the variety of goals selected by colleges of business and management will generate a similar variety in the topics of manuscripts submitted. Additionally, we invite manuscripts focused on curriculum, scholarship, and other initiatives and are particularly interested in manuscripts that describe and evaluate initiates related to external stakeholders. Finally, while the journal has a regional mission, we invite publications with a local, regional, national, and international scope.

Suggested topics include, but are not limited to, the following:

Eliminating the Digital Divide. Although the digital divide was identified some time ago and has been studied and addressed over time, discrepancies persist and became increasingly problematic during the COVID-19 pandemic (Lythreatis, 2022). We welcome manuscripts addressing the digital divide with a focus on studies and initiatives designed to work toward an elimination of the digital divide regionally, nationally, and globally. Initiatives designed to eliminate the digital divide have the potential to reduce poverty, improve health, promote quality education, improve gender equality, improve economic growth in poorly served areas, reduce inequalities, and boost industry, innovation, and infrastructure (United Nations, 2024).

Use of Information Systems to Support the Development of Resilient Agriculture. The resilience of farming and agriculture is a pressing issue in the Midwest region of the United States. Resilient agriculture is the philosophy that land management methods should preserve the potential of agricultural land for future production and economic gain. Additionally, the ability of the agricultural sector to adapt and continue to innovate is essential (Center for Resilience, 2024). As part of ensuring the resilience of the agricultural sector, there is a clear need for farms in the Midwest region of the United States and beyond to become better prepared for the challenges of weather extremes and climate variability (Williams et al., 2022). Less clear is the role that information systems may play in building agricultural resilience. Access to climate information has been proposed as an essential feature of agricultural resilience; however, challenges in the use of this information have been found and future research and initiatives can be designed to better understand and address these challenges (Chaudhuri & Kendall, 2021; Haworth et al., 2018; Savari et al., 2024).

Information systems faculty working in the Midwest region of the United States are well positioned to undertake research and community initiatives focused on the potential benefits of and barriers to the use of information systems in the development of resilient agriculture. Additionally, faculty may partner with colleagues in agricultural colleges to develop curriculum at the intersection of information systems and agricultural innovation. The *Journal of the Midwest Association for Information Systems* invites publications focused on these initiatives in support of the zero hunger and climate action goals of the United Nations Sustainable Development Goals (United Nations, 2024).

Use of Information Systems to Monitor and Improve Water Quality. Polyfluorinated substances (PFAS) have been found in many public water systems and private wells nationally, and recent EPA regulations require the removal of these chemicals (Friedman, 2024). E. Coli, nitrates, and phosphorus also threaten the quality of water in the Midwest (Schneider, 2023). The role of information systems in monitoring and improving water quality will be an essential part of any initiatives designed to improve water quality (Behmel et al., 2016), and the journal welcomes manuscripts addressing these issues which are linked to the United Nations goals of clean water and health (United Nations, 2024).

High Impact Teaching Practices and Information Systems Curriculum Innovations. The *Journal of the Midwest Association for Information Systems* has been and will continue to be an outlet for the publication of manuscripts on high impact teaching practices and curricular innovations in the field of information systems. Work on these topics supports the United Nations Sustainable Development Goals of quality education, decent work, and economic growth (United Nations, 2024).

Gender Equality in Information Systems Education and Practice. The *Journal of the Midwest Association for Information Systems* has also been an outlet focused on gender equality in information systems education and practice as seen in Table 2. The journal continues to welcome manuscripts in these areas which support the United Nations Sustainable Development Goals of gender equality and reduced inequality (United Nations, 2024). Manuscripts focused on gender differences in access to and use of information systems (e.g., Shah & Krishnan, 2023), gender equity and artificial intelligence (e.g., Paton-Romero et al., 2022), and gender equity and blockchain technology (e.g., Di Vaio et al., 2023) as well as those focused on other information systems topics are encouraged.

Remote and Hybrid Work. The *Journal of the Midwest Association for Information Systems* encourages the submission of manuscripts on remote and hybrid work arrangements which may build on the work of Hadidi and Klein (2024). Work in this area is consistent with accreditation emphasis on societal impact by supporting the United Nations Sustainable Development Goals of eliminating poverty, decent work, and economic growth (United Nations, 2024).

4. Conclusion

As accreditation standards place more emphasis on societal impact, information systems faculty may examine their

research agendas, curricular focus, and other initiatives to find and develop ways in which their work may have a positive societal impact on their local, regional, national, and international communities. Potential initiatives may be found in the United Nations Sustainable Development Goals taxonomy (United Nations, 2024) or through other frameworks that may be adopted by their colleges of business and management. While accreditation standards do not demand that all faculty have a focus on societal impact (AACSB, February 2023), over time, many information systems faculty may become interested in such a focus. As work with an emphasis on societal impact increases in support of various accreditation standards, the *Journal of the Midwest Association for Information Systems* encourages submission of manuscripts on curriculum, scholarship, and other initiatives so that ideas can be shared across the Midwest and beyond as faculty build expertise and collaborate to work more intentionally in these areas.

5. Overview of the Contents of this issue

This issue of the journal includes two traditional research articles. Troy Strader and Royce Fichtner in their interesting, and curriculum related article looked at 163 US universities to see the extent of coverage of the Information Technology Ethical and Legal (ITEL) coverage in their curriculum. At this age of Generative AI, their findings and set of recommendations is very timely for IS and IT professionals to read about.

Austin Coursey, Matthew Tennyson, and Vlad Krotov in their important and informative article looked at the authorship attribution related to the R code. Given the popularity of R code these days, they propose a tool to properly attribute the generated R codes to the original author(s).

We appreciate and wish to acknowledge the contributions of reviewers for this issue of the journal, including Queen Booker (Metropolitan State University), Mari Buche (Michigan Technological University), Yi “Maggie" Guo (University of Michigan-Dearborn), Bryan Hosack, (Penske Logistic), and Alanah Mitchell (Drake University).

6. References

AACSB. (2023). Guiding principles and standards for business accreditation. https://www.aacsb.edu//media/documents/accreditation/2020-aacsb-business-accreditation-standards-june-2023.pdf?rev=d31cfbe864e54792816ff426fe913e65&hash=33A159779F107443A64BDACBBB7000C5.

AACSB. (February 2023). AACSB and societal impact: Aligning with the AACSB 2020 business accreditation standards. https://www.aacsb.edu/-/media/documents/accreditation/aacsb-and-societal-impact.pdf?rev=8b09ad970fb6445b9327c91f3fea5708&hash=793FE4D886040B6C499B1FC12B7E3835.

ABET. (2024). Criteria for accrediting computing programs, 2024-2025. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-computing-programs-2024-2025/

ACBSP. (2024). Accreditation Council for Business Schools and Programs Standards, Version C, February 2024. https://cdn.ymaws.com/acbsp.org/resource/resmgr/docs/accreditation/Unified_Standards_and_Criter.pdf

Behmel, S., Damour, M., Ludwig, R., and Rodriguez, M.J. (2016). Water quality monitoring strategies - A review and future perspectives. *Science of the Total Environment*, 571, 1312-1329.

Center for Resilience in Agricultural Working Landscapes. (2024). https://centerforresilience.unl.edu/agricultural-resilience#:~:text=Agricultural%20resilience%20ensures%20that%20management,and%20adapt%20when%20shocks%20occur. accessed May 18, 2024.

Chaudhuri, B., and Kendall, L. (2021). Collaboration without consensus: Building resilience is sustainable agriculture through ICTs. *The Information Society*. 37(1), 1-19.

Chen, A.J.W., Boudreau, M., and Watson, R.T. (2008). Information systems and ecological sustainability. *Journal of Systems and Information Technology*. 10(3), 186-201.

Clemons, E. K., Dewan, R. M., Kauffman, R. J., and Weber, T. A. (2016). Special section: When machine meets society: Social impacts of information and information economics. *Journal of Management Information Systems*, *33*(2), 542–545.

De Leoz, G., Petter, S., Peffers, K., Tuunanen, T., and Niehaves, B. (2018). Considering the social impacts of artifacts in information systems design science research. *European Journal of Information Systems*, *27*(2), 154–170.

Di Vaio, A., Hassan, R., and Palladino, R. (2023). Blockchain technology and gender equality: A systematic literature review. *International Journal of Information Management*, 68, 102517. https://doi.org/10.1016/j.ijinfomgt.2022.102517.

Dwivedi, Y.K., Kshetri, N., and Hughes, L. (2023). Exploring the darkverse: A multi-perspective analysis of the negative societal impacts of the metaverse. *Information Systems Frontiers*, 25, 2071-2114.

Eierman, M. A., and Iversen, J. (2018). Comparing test-driven development and pair programming to improve the learning of programming languages. *Journal of the Midwest Association for Information Systems*. 2018(1), Article 3.

Fernández, E. (2017). Innovation in healthcare: Harnessing new technologies, *Journal of the Midwest Association for Information Systems*. 2017(2), Article 8.

Friedman, (2024). E.P.A. says 'forever chemicals' must be removed from tap water. *The New York Times*. (April 10, 2024). https://www.nytimes.com/2024/04/10/climate/epa-pfas-drinking-water.html.

George, J. F., and Hadidi, R. (2022). COVID-19 and its impact on the Midwest United States. *Journal of the Midwest Association for Information Systems*. 2022(1), Article 1.

Hadidi, R., and George, J. F. (2022). COVID-19 and examples of "best" teaching practices from the lens of different stakeholders. *Journal of the Midwest Association for Information Systems*. 2022(2), Article 1.

Hadidi, R., and George, J. F. (2023). Potential uses of AI-based platforms in teaching and learning. *Journal of the Midwest Association for Information Systems*. 2023(2), Article 1.

Hadidi, R., and Klein, B. D. (2024). The future of work, physical location of workers, technological issues and implications. *Journal of the Midwest Association for Information Systems*. 2024(1), Article 1.

Hadidi, R., Power, D., and George, J. F. (2016). Information technology is transforming the heartland: Making the case for Midwest United States. *Journal of the Midwest Association for Information Systems*. 2016(1), Article 1.

Haworth, B.T., Biggs, E, Duncan, J., Wales, N., Boruff, B., and Bruce, E. (2018). Geographic information and communication technologies for supporting smallholder agriculture and climate resilience. *Climate*. 6(4), 97.

Heavin, C. (2017). Health information systems – Opportunities and challenges in a global health ecosystem. *Journal of the Midwest Association for Information Systems*. 2017(2), Article 1.

Kenny, G., O' Connor, Y., Eze, E., and Heavin, C. (2017). Trends, findings, and opportunities: An archival review of health information systems research in Nigeria. *Journal of the Midwest Association for Information Systems*. 2017(2), Article 6.

Kiely, G. L., Heavin, C., and Lynch, P. (2019). Building a shared understanding of female participation in IT through collaboration: A shared mental model approach. *Journal of the Midwest Association for Information Systems*. 2019(1), Article 3.

Lebens, M. (2021). Using prototyping to teach the design thinking process in an asynchronous online course. *Journal of the Midwest Association for Information Systems*. 2021(2), Article 3.

Luse, A., and Burkman, J. (2018). Safely using real-world data for teaching statistics: A comparison of student performance and perceived realism between dataset types. *Journal of the Midwest Association for Information Systems*. 2018(1), Article 2.

Lythreatis, S., Singh, S.K., and El-Kassar, A.-N. (2022). The digital divide: A review and future research agenda. *Technological Forecasting and Social Change*. 175, 121359.

Malhotra, A., Melville, N.P., and Watson, R. T. (2013). Spurring impactful research on information systems for environmental sustainability. *MIS Quarterly*, 37(4), 1265-1274.

Marett, K., and Xiao, S. (2022). Broadband Internet access as a localized resource for facilitating information security knowledge. *Journal of the Midwest Association for Information Systems*. 2022(1), Article 2.

Melville, N.P. (2010). Information systems innovation for environmental sustainability. *MIS Quarterly*, 34(1), 1-21.

Mitchell, A. J. D. (2018). Small business website development: Enhancing the student experience through community-based service learning. *Journal of the Midwest Association for Information Systems*. 2018(2) Article 4.

Muraski, J. M. (2023). IS career day in a class: Raising college student awareness and interest in information systems. *Journal of the Midwest Association for Information Systems*. 2023(1), Article 3.

Muraski, J.M., and Iversen, J. (2022). Growing computer science and information technology education in K-12: Industry demand and ecosystem support. *Journal of the Midwest Association for Information Systems*. 2022(2), Article 2.

Muraski, J.M., Iversen, J., and Iversen, K.J. (2021). Building collaboration networks and alliances to solve the IT talent shortage: A revelatory case study. *Journal of the Midwest Association for Information Systems*. 2021(1), Article 3.

Paton-Romero, J.D., Vinuesa, R., Jaccheri, L., and Baldassarre, M.T. (2022). State of gender equality in and by artificial intelligence. *International Journal on Computer Science and Information Systems*. 17(2), 31-48.

Power, D., and Hadidi, R. (2019). Transforming agriculture: Exploring precision farming research needs. *Journal of the Midwest Association for Information Systems*. 2019(2), Article 1.

Rowland, P., and Noteboom, C. B. (2018). Anchoring female millennial students in an IT career path: The CLASS anchor model. *Journal of the Midwest Association for Information Systems*. 2018(2), Article 3.

Sabherwal, R., and Grover, V. (2024). The societal impacts of generative artificial intelligence: A balanced perspective. *Journal of the Association for Information Systems*, 25(1), 13-22.

Savari, M., Zhoolideh, M., & Limuie, M. (2024). An analysis of the barriers to using climate information services to build a resilient agricultural system in Iran. *Natural Hazards*, 120, 1395-1419.

Schneider, K. (2023). New U.S. climate law could make Midwest water contamination worse. (February 10, 2023). https://www.greatlakesnow.org/2023/02/new-u-s-climate-law-could-make-midwest-water-contamination-worse/ . accessed May 18, 2024.

Shah, C.S., and Krishnan, S. (2023). Digital gender gap, gender equality and national institutional freedom: A dynamic panel analysis. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-023-10456-9.

Strader, T. J., and Bryant, A. (2018). University opportunities, abilities and motivations to create data analytics programs. *Journal of the Midwest Association for Information Systems*. 2018(1), Article 4.

United Nations Department of Economic and Social Affairs. (2024). Sustainable Development Goals. https://sdgs.un.org/goals, accessed May 19, 2024.

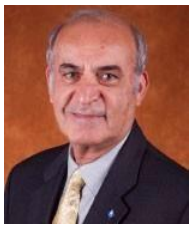
Vroegindeweij, R., and Carvalho, A. (2019). Do healthcare workers need cognitive computing technologies? A qualitative study involving IBM Watson and Dutch professionals. *Journal of the Midwest Association for Information Systems*. 2019(1), Article 4.

Williams, T., Schmitz, H., and Shulski, M. (2022). Resilient agriculture: Weather ready farms. https://www.climatehubs.usda.gov/sites/default/files/Weather_Ready_eFieldbook.pdf . (accessed May 18, 2024).

Witt, C., Melton, J., and Miller, R. E. (2024). Gender, emotional intelligence, and the need for popularity: Exploring the causes of faux pas posting beyond the behavior of friends. *Journal of the Midwest Association for Information Systems*. 2024(1), Article 3.

Author Biographies

Barbara D. Klein is Professor of Management Information Systems and Information Technology Management at the University of Michigan-Dearborn. She received her Ph.D. in Information and Decision Sciences from the University of Minnesota, her M.B.A. from the State University of New York at Albany, and her B.A. from the University of Iowa. Professor Klein has published in the *Journal of the Midwest Association for Information Systems*, *MIS Quarterly*, *Omega*, *Database*, *Information & Management*, *Information Resources Management Journal*, and other journals. Her research interests include information quality, user error behavior, and information systems pedagogy. Professor Klein has also worked in the information systems field at IBM, Exxon, and AMP.



Rassule Hadidi is Dean of the College of Business and Management, Metro State University, Minneapolis, Minnesota. His current research areas of interest include online and blended teaching and learning pedagogy and its comparison with face-to-face teaching; curriculum development and quality assessment; cloud computing and its applications for small and medium-sized enterprises; and quality of online information. He has served as the president as well as the At-Large Director of the Midwest Association for Information Systems and is the founding Managing Editor of the *Journal of the Midwest Association for Information Systems*. He is an AIS Distinguished Member – Cum Laude and is a member of the Board of Directors of the Society for Advancement of Management.

This page intentionally left blank.

Journal of the Midwest Association for Information Systems

Volume 2024 | Issue 2 Article 2

Date: 07-02-2024

IT Ethics and Law Courses: An Analysis of Curricula in Medium-Sized US Doctoral Classification Universities

Troy J. Strader
Drake University, troy.strader@drake.edu

J. Royce Fichtner
Drake University, royce.fichtner@drake.edu

Abstract

Information technology ethical and legal (ITEL) issues must be incorporated into university curricula to properly train new information technology (IT) specialists to deal with the full range of issues they will face in their careers. This report looks at the extent to which ITEL classes are being developed and taught in a sample of US universities. The study identifies which academic departments offer the courses, which university resource, governance, and enrollment profiles are most often associated with the offering of ITEL courses, and what topics are most commonly included in these courses. The sample of schools reviewed are the 163 US universities that are in the Carnegie classification for four-year medium-sized doctoral universities. Findings show that about five out of every eight universities that were reviewed offer at least one ITEL class and the largest number of unique ITEL courses at any university is five. The courses are most commonly taught at the undergraduate level in computer science. Further analysis shows that the universities that offer at least one ITEL class have larger total student enrollments, are publicly governed, and have a majority, or larger proportion, of undergraduate students. Course descriptions typically discuss the topics in general terms referring to the broad subjects of IT-related ethics and law, but it should be expected that AI's ethical and legal issues will receive more attention in the future. The report concludes with a set of recommendations for universities that are developing new ITEL courses and directions for future research.

Keywords: information technology ethics, information technology law, information systems curricula

DOI: 10.17705/3jmwa.000089

1. Introduction

In the decades after the invention of the computer, opportunities for new applications were most often determined by the extent to which processor speed and storage capacity had increased. In today's environment, it is not so much a matter of what can be done with information technology, but rather what should be done. Ethical and legal issues are much more significant. This is reflected in industry practice, but they are also an important consideration when developing new curricula. Information technology ethical and legal (ITEL) issues must be incorporated into university curricula to properly train new information technology specialists so they are prepared to deal with the issues they will face in their careers. Review task forces have recommended that this topic be incorporated into both undergraduate and graduate information systems (IS) curricula. One of the nine information systems competency areas included in the MSIS 2016 master's degree global competency model is ethics, impacts, and sustainability (Topi *et al.*, 2017). The task force's rationale for adding this competency is that MSIS 2016 focuses on capabilities that enable graduates to contribute to positive transformation of various societal activities through digitalization. They define ethics, impacts, and sustainability as the conceptualization and implementation of environmentally and socially sustainable IT solutions that are aligned with the responsibilities of organizations and in compliance with legislative and regulatory requirements and industry standards. The ACM/AIS IS2020 undergraduate IS curriculum review task force also recommended the addition of an organizational domain competency for ethics, use and implications for society (Leidig and Salmela, 2022). Their rationale for adding this competency is that it is related to the ubiquitous nature of information systems, and increasing influence of these systems in society.

These recommendations were made several years ago so there has been sufficient time for programs to incorporate these topics into their undergraduate majors and graduate degree programs. Given how important it is for industry to ethically and legally develop, implement, and operate global information systems, the purpose for this report is to evaluate the current educational environment and assess the extent to which ITEL classes are being developed and taught in US universities and, more specifically, which courses have been developed. The following five questions will be addressed in this study:

1. What portion of the reviewed universities offer one, or more, ITEL classes?
2. At what level (undergraduate vs. graduate) are the courses offered?
3. Given that these courses are interdisciplinary, which departments most often offer these ITEL classes?
4. What are the resources, governance structures, and enrollment profiles for universities that are more likely to offer ITEL courses?
5. What topics are most often included in ITEL courses?

The answers to these questions will be based on a review of a sample of universities that represent medium-sized US universities located in all regions of the country. This provides a sample of universities that includes both public and private universities that would be large enough to have the intellectual and financial capacity to offer ITEL courses.

The following sections are included in this report. First, a review of related literature is provided to describe similar IT-related curriculum studies and the questions they have addressed. This is followed by a description of the methodology used in this study, the research questions, and the answers that were found based on the analysis of the course and university data that was collected and reviewed. The report concludes with a set of recommendations for universities that are developing new ITEL courses, study limitations, and directions for future research.

2. Curriculum Review Studies

A number of past studies have reviewed, or recommended designs for, IT-related curricula. Some of the studies make recommendations for overall curriculum design while others make recommendations for changes that could be made within existing courses. An additional study looked at the university characteristics that support more up-to-date and innovative IT curricula.

2.1 Studies Recommending Overall Curriculum Change

Gupta et al. (2015) presented a model curriculum that introduces business intelligence (BI) and analytics topics into existing curriculum. It focused on adding appropriate elective courses to existing curriculum in order to foster the development of BI skills, knowledge, and experience for undergraduate majors, master of science in business information systems degree students, and MBAs.

Burns et al. (2018) investigated the knowledge and skills required by potential employers of students graduating from undergraduate information systems programs. Entry level job listings were collected and analyzed from several Internet sites specializing in technology related employment. This information was used as the basis to compare the knowledge and skills required by potential employers to the suggested curriculum of the 2010 ACM/AIS Information Systems Curriculum Guidelines.

Saltz et al. (2018) explored the different data science codes of conduct and ethics frameworks. They compared this analysis with the results of a systematic literature review focusing on ethics in data science. Their analysis identified twelve key ethics areas that should be included within a data science ethics curriculum. The results from their study can be used by educators and program coordinators to identify key ethical concepts that can be introduced within a data science program.

Lyytinen et al. (2021) describe the first phase of the work done by the Management Curriculum for the Digital Era (MaCuDE) disciplinary task force on information systems. The MaCuDE project recommends changes to business curricula based on the influence of digital technologies on business transformation and the widespread use of big data analytics (BDA) and AI technologies in organizations. Based on the MaCuDE project survey conducted in early 2020, they identified the core digital topics and tools that the programs covered based on a sample of global IS departments from 17 undergraduate programs and 23 graduate programs.

Another study used text mining techniques to analyze university information systems curricula (Föll and Thiesse, 2021). It presented a quantitative content analysis procedure for collecting, analyzing, evaluating, and comparing curricula that provides an alternative to qualitative content analysis. The procedure was tested using data from more than 90 German IS programs and the results provided insights for curriculum redesign and assessing whether programs fulfill the skill expectations that employers have for their new IT employees.

2.2 Studies Recommending Changes for IT Courses

One of the earliest studies in this area recommended adding a course on information technology law for legal education in the US (Hirsh and Miller, 2003). They noted that legal education in the United States has been fundamentally unchanged in the past century while the practice of law has been revolutionized by information technology. The authors reviewed the availability of courses covering use of technology in law practice at American law schools and set out their own proposal for such a course at the Duke University School of Law. They focused on two areas – technology use in the courtroom, and technology use in legal offices. At the time, they were concerned with the technology skills required to participate in the legal profession. Course content related to laws governing data management and privacy would be a consideration in the future.

Another study outlined a series of ten themes for teaching ‘cyberlaw’ in an attempt to overcome problems associated with teaching a course on IT law that is new and rapidly changing (Quirk, 2008). The themes identified include jurisdiction, agency, payments, risk transfer, security, taxation, crime, history, privacy and intellectual property. The article discusses each theme in relation to the legal environment at the time of the study along with advanced sub-topics that may be relevant in the future.

Subramanian and White (2008) discuss the evolution of information technology and how this has led to a plethora of US and international laws that govern the use of IT. Given the importance of these laws to IT managers, they review model IT curricula and found that legal issues were not receiving the attention that they deserved. Their study provides the justification and design for a course in IT and the Law.

Grosz et al. (2019) noted how important it is that computer science curricula expand to include ethical reasoning about the societal value and impact of information technologies. In their study, they describe Embedded EthiCS, a novel

approach to integrating ethics into computer science education that embeds philosophers teaching ethical reasoning directly into computer science courses.

Fiesler et al. (2020) described current trends in computing ethics coursework by conducting a qualitative analysis of 115 syllabi from university technology ethics courses. They identified the content and goals for these courses and made recommendations for how these courses might be integrated across a computing curriculum.

2.3 A Study Identifying the University Characteristics that Support New IT-Related Curriculum Development

Strader and Bryant (2018) conducted a study to identify the characteristics of schools that have developed data analytics programs. The study identified factors that increase the likelihood that a university will develop a data analytics program based on a review of 391 US regional master's universities. The study found that schools with data analytics programs are more likely to be in larger cities and have larger student enrollments, better educational quality rankings, and existing statistics and/or actuarial science programs.

While past studies have addressed a variety of IT curriculum and course issues indicating where changes were needed, no study has assessed the current environment for ITEL course offerings and the topics that are included. This present study looks at this important and timely IT curriculum issue.

3. Methodology

The sample of universities reviewed in this study are in the Carnegie classification that includes medium sized doctoral universities (Carnegie Classification of Institutions of Higher Education, 2023). This is a representative sample of US universities that could provide some insight into current ITEL course development. Medium sized universities were chosen for the sample because very small universities are unlikely to offer these courses because they don't have the financial or faculty resources, and curriculum at very large universities would not be representative of the much larger number of mid-sized universities.

There are 163 universities in this group. Data for each of the universities was collected from a variety of sources. A spreadsheet was downloaded from the Carnegie website that included data for all of the institutions that were Doctoral Universities for their Basic Classification. Within this spreadsheet, the universities were selected when their Size & Setting included the word "medium" indicating that they were a medium sized university relative to the size of all of the other doctoral universities. This spreadsheet also included more specific data about each university's governance (public versus private) and enrollment profile (ranging from very high undergraduate through very high graduate). Each of the university websites were then searched to find their course catalog. The course catalogs were manually examined to identify courses that were primarily related to the topic of interest in this study – information technology ethics and law. Typically, the courses that matched this topic were found in business information systems, computer science, business/data analytics, philosophy, business law, or related areas. For each course, data was collected for the name of the department offering the course, the course number, title and description, and the level at which the course was offered (undergraduate and/or graduate). Some universities did not offer any relevant courses, while others offered more than one. Additional university data was then added for the overall enrollment, and whether they had undergraduate majors or master's programs in business information systems, computer science, business/data analytics, or philosophy. Total enrollment data was found in the National Center for Educational Statistics, COLLEGE Navigator, website (National Center for Educational Statistics, COLLEGE Navigator, 2023). This site was used for enrollment data instead of searching each university website because the data was more consistent and it included both full-time and part-time student numbers.

The combined set of collected data was used in this study to assess the current state of ITEL course offerings from several perspectives. The data look at what courses are offered, their departments and levels, the resource, governance, and enrollment profile strategic characteristics of universities who are more likely to offer these courses, and the course topics. The specific questions and findings are discussed in the next section.

4. Questions and Findings

4.1 Portion of Universities Offering ITEL Courses

The first question addressed in this study was to identify what portion of the 163 reviewed universities offered at least one ITEL course? It was found that 102 (62.6%) of the universities offered at least one ITEL course. In addition, 44 (27.0%) of the universities offered more than one ITEL course. These findings point to the current diffusion stage for the development of these courses. Fichman (1992) found that the diffusion process usually starts out slowly among pioneering adopters, reaches "take-off" as a growing community of adopters is established, and levels-off as the population of potential adopters becomes exhausted. This can be viewed as an S-shaped three-phase cumulative adoption curve. Given this pattern, the findings from this study indicate that ITEL courses are several years into a diffusion process where it is past the early adopter phase and well into the second phase where a larger number of universities have offered these courses for several years. The number of new ITEL courses offered in the next decade can be expected to be relatively small. Another way to view the course data is to look at the distribution of the number of ITEL courses offered at each of the reviewed universities. For each of the universities reviewed, how many different ITEL courses do they offer? The results are summarized in Figure 1.

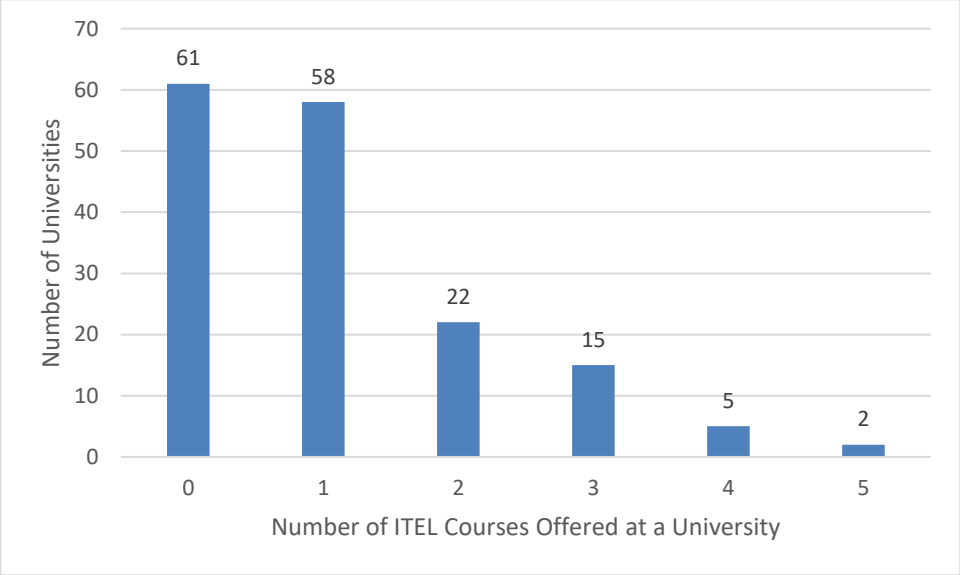


Figure 1. Distribution of How Many Different ITEL Courses are Offered at How Many Universities

As discussed earlier, 61 of the 163 universities do not offer an ITEL course. For the universities that do offer an ITEL course, most of them (58 out of 102, 56.9%) only offer one course. The largest number of unique ITEL courses offered is five. These findings would be expected for a new course offering that can cover a wide range of IT ethical and legal issues at either the undergraduate or graduate levels. Two or more courses would typically indicate that the courses are offered at different levels (undergraduate vs. graduate), or offered by different departments that each offer a unique focus on the issues.

4.2 ITEL Course Level Distribution (Undergraduate vs. Graduate)

The second question addressed in this study looked at the level in which these ITEL courses are offered. What portion of the ITEL courses offered were at the undergraduate level, graduate level, or a cross-listed course for both undergraduate and graduate students? The results are summarized in Figure 2. The majority of the ITEL courses are at the undergraduate level (119 out of 177, 67.2%). The remaining courses are either graduate level (48, 27.1%), or unique courses that are cross-listed courses for both undergraduate and graduate students (10, 5.6%). It is likely that the larger number of undergraduate courses is because the universities offer more undergraduate programs and have a relatively larger undergraduate student population when compared with their graduate student enrollment. It does not indicate that the topic is less relevant for graduate students. For some universities, offering a cross-listed course at both levels provides an efficient method to deliver the content to a broader audience.

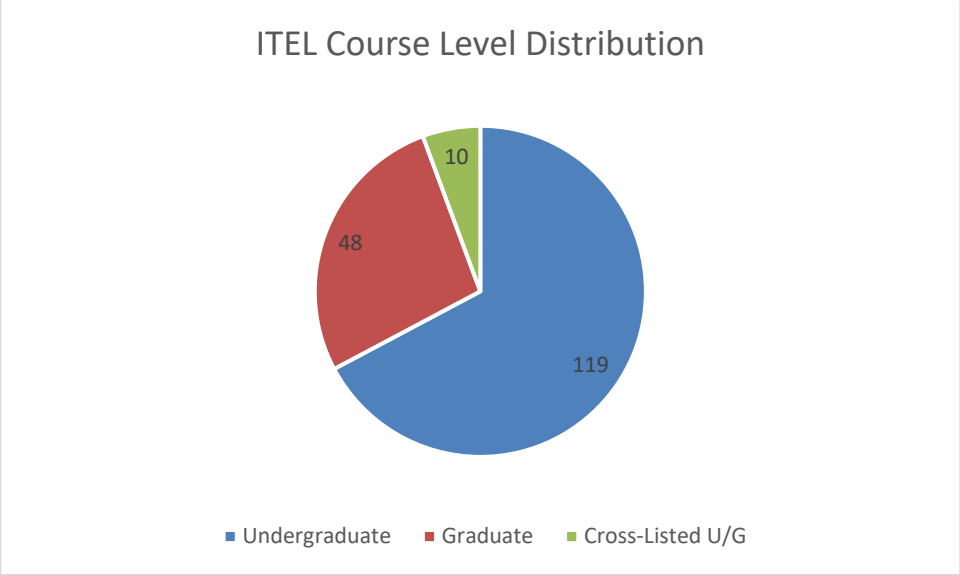


Figure 2. Distribution of ITEL Courses Offered at Undergraduate and/or Graduate Levels

4.3 Departments Offering ITEL Courses

The third study question looks at which departments are most likely to offer at least one ITEL course. The results are summarized in Table 1. As expected, the most common department offering an ITEL class is computer science. Ethical and legal issues related to information technology development and impact are important components of a computer science education in today's world. Philosophy is the second most common department offering ITEL classes. Ethics is a major topic in any philosophy curriculum. The explosive growth and impact of information technology make it an obvious area for discussion and curriculum inclusion. The third and fourth most common departments offering ITEL classes are business information systems and business/data analytics. They are the areas that collect, store, distribute and analyze massive amounts of data so ethics and relevant data-related laws are an important topic to include in their curricula to properly train students to work in their field.

Department	Number Offering ITEL Course	Percentage Offering ITEL Course
Computer Science	79	44.6%
Philosophy	30	17.0%
Business IS/MIS	28	15.8%
Data/Business Analytics	19	10.7%
Business Law	8	4.5%
Other	5	2.8%
Business	4	2.3%
Multi-department	4	2.3%

Table 1. Distribution of Departments Offering ITEL Courses

4.4 University Characteristics for Schools Offering ITEL Courses

To answer the fourth question addressed in this study, an additional consideration when assessing ITEL course offerings is to identify the underlying characteristics of universities that are most likely to offer at least one ITEL course. In this section the focus is on three university characteristics: (1) total university enrollment, (2) how the programs are governed, and (3) the university's enrollment profile strategy. These characteristics may provide the underlying resources and motivations to create ITEL courses as part of an innovative and up-to-date curricula.

4.4.1 University Total Enrollment

Universities with larger total student enrollments would be more likely to have more overall university financial resources. It would be expected that these resources would provide a greater opportunity to offer at least one ITEL course. A regression model was used to test this idea. The dependent variable is a binary variable indicating whether, or not, one or more ITEL courses are offered (0=no ITEL course, 1=one or more ITEL courses). The independent variable is the university's total student enrollment. The smallest university has 1014 students, the largest has 18053, and the average total enrollment is 7667. The analysis supports this expected relationship. The results shown in Table 2 indicate that universities with larger total enrollments are significantly more likely to offer at least one ITEL course when compared with universities with smaller enrollments.

Parameter	Estimate (S.E.)	p-value
Intercept	0.4520 (0.1021)	<0.00001***
Total Univ. Enrollment	0.00002 (0.00001)	0.00007***

NOTE: ***p < .01; **p < .05; *p < .10

Table 2. Relationship Between Total University Enrollment and Offering an ITEL Course

4.4.2 Governance

Universities may be governed as public universities or private universities. One major difference is that public universities would have some level of governmental support that would provide additional financial resources relative to their private university counterparts. The number of publicly and privately governed universities that offer, or do not offer, at least one ITEL course is shown in Table 3. The data supports the idea that publicly governed universities are more likely to offer at least one ITEL course. 80.4% of public universities in the sample offer at least one ITEL course, while only 53.3% of private universities offer an ITEL course.

University Governance	Offer ITEL Course	Do Not Offer ITEL Course	Percentage Offering ITEL Course
Public	45	11	80.4%
Private	57	50	53.3%

NOTE: There are 56 publicly governed universities and 107 privately governed universities in the sample.

Table 3. Percentage of Publicly and Privately Governed Universities Offering an ITEL Course

4.4.3 Enrollment Profile Strategy

Universities may choose to focus their efforts primarily on undergraduate education, or they may decide to have a greater portion of graduate students. The data from the Carnegie website separates universities into one of four categories: (1) very high undergraduate, (2) high undergraduate, (3) majority undergraduate, or (4) very high graduate. Table 4 shows the number of universities in each category that offer, or do not offer, at least one ITEL course. It appears that undergraduate focused universities are far more likely to offer an ITEL course. This matches the results shown earlier in Figure 2. 81.8% of universities with a very high undergraduate enrollment profile strategy offer an ITEL course, while only 37.5% of universities with a very high graduate enrollment profile strategy offer a course.

Enrollment Profile Strategy	Offer ITEL Course	Do Not Offer ITEL Course	Percentage Offering ITEL Course
Very High Undergraduate	9	2	81.8%
High Undergraduate	49	25	66.2%
Majority Undergraduate	38	24	61.3%
Very High Graduate	6	10	37.5%

Table 4. Enrollment Profile Strategies for Universities Offering an ITEL Course

The results from the analysis described above shows that the universities that offer at least one ITEL class have larger total student enrollments, are governed as a public university, and have a much larger proportion of undergraduate students. The findings show that resources are an important enabler for new course development. More students and public governance provide more financial resources which provides greater opportunities for new course development.

4.5 Course Topics

For the fifth study question, the final perspective that can provide insights into the current state of ITEL classes is to identify which topics are included in the courses. One of the most popular textbooks in this area is organized into ten chapters that builds from broad background discussions of ethics and law and then identifies several of the most impacted topic areas such as privacy, security, and intellectual property (Reynolds, 2019). Each course description was reviewed and the number of times each of these major topics appeared is summarized in Table 5. This provides a list of topics that are most emphasized in ITEL courses, which ones appear less often, and which topics do not appear in any of the course descriptions. Results are displayed in three columns – all courses, only undergraduate courses, and only graduate courses.

There are five broad topic areas that appear 100 or more times. As expected, ethics/ethical and law/legal appear most often in the course descriptions because they are very broad terms that are directly related to the course topic. Other terms commonly included are data, privacy, and some mention of security. These are some of the most important areas where laws exist to protect individuals and organizations along with opportunities for unethical behavior. Seven additional topics appear more than ten times. This is where the descriptions point to areas of emphasis for a particular course. These more specific topic areas include ethical and legal issues for intellectual property, the Internet, artificial intelligence, the economy, professional codes of ethics (conduct), software development, and social media/networks. Finally, topics such as social responsibility, labor and productivity, social audit, freedom of expression, and outsourcing rarely appear in the course descriptions, if at all. Overall, it appears that most course descriptions are written in broad terms so they can evolve as specific ethical contexts and laws change over time. A smaller number are more narrowly defined to fit a more specific purpose within a curriculum. It is very rare that specific ethical frameworks or laws are noted in a course description.

There are a few differences that are apparent when comparing undergraduate and graduate course descriptions. Undergraduate course descriptions mention ethics/ethical more often while graduate courses mention law/legal and data relatively more often. Surprisingly, artificial intelligence is not included many times in graduate course descriptions, but that could be the case where the course description was written several years ago before AI began to receive increased attention. This is the topic that would most likely see greater interest going forward.

Rank	Word/Phrase	Number of Appearances for Word/Phrase in a Course Description		
		All Courses (Undergrad and Grad + U/G cross-listed)	Only Undergrad	Only Grad
1	Ethics/ethical	274	210	46
2	Law/legal	196	97	78
3	Data	142	65	70
4	Privacy	120	80	40
5	Security/cybersecurity	100	57	37
6	Intellectual property, copyright, patent, trademark	62	42	18
7	Internet	32	25	7
8	Artificial intelligence/AI, intelligent system	30	23	5
9	Economy/economics	18	11	5
10	Professional code of ethics	16	13	2
11	Software development/engineering	14	12	0
12	Social media/network	12	9	1
13	Social responsibility	7	5	1
14	Labor, productivity	2	2	0
NR	Social audit Freedom of expression, first amendment Outsourcing	0	0	0

Table 5. Number of Times a Word/Phrase Appears in a Course Description

5. Recommendations for New ITEL Courses

The findings from this study, and the undergraduate and graduate information systems curriculum task force reports, provide some recommendations that universities can consider when developing new ITEL courses. The goal is to develop a course, or courses, that effectively teaches students about the current legal environment, ethical decision making, and industry practice, but does it in a way that efficiently utilizes available financial and faculty resources. These recommendations pertain to: (1) who should offer the course, (2) what topics should be included in general or program-specific ITEL undergraduate or graduate courses., and (3) a set of complementary assessment methods.

An important consideration when developing new ITEL courses is to try to utilize existing faculty capacity so that the new course does not require excessive resources. The most efficient way to develop a first ITEL course is to create one course that is cross-listed, or offered as an elective, between two or more departments. Based on an analysis of existing courses, the most common combination of departments is computer science paired with philosophy, business information systems, or business/data analytics. One general ITEL course can serve students in all of these programs. This is the most efficient way to teach students about these topics without requiring large amounts of new resources. The course can be required for all four of these majors because the content is so important in today's technological world. If it is required, then more students will take this course leading to larger class sizes that demonstrate that the course is needed and should be continued to be offered in the future. The same recommendation can be made for a graduate level ITEL class. It can be cross-listed for master's programs in information systems, computer science, business administration, accounting, public administration, and philosophy. Unique versions could also be developed to focus on specific issues that are unique to law schools or healthcare programs. Many of the courses assessed in this study did not require any prerequisite courses, but it may be helpful to require an introduction to information systems and/or an introduction to business law course to reduce the time required to cover these background issues.

The second major consideration is course content in new ITEL classes. Table 6 provides one example describing how to organize topics for a general undergraduate or graduate ITEL class. It lists topics and the rationale for their inclusion. The courses can be broken into three sections: (1) introduction to ethical analysis and the law, (2) analysis of ethical and legal issues for specific IT-related topics and contexts, and (3) comprehensive analysis of ethical and legal issues for an organization's entire information system or e-commerce site looking at the interrelationship between issues and potential conflicts. The ACM/AIS IS2020 competency model and the MSIS 2016 report both recommend the addition of an ITEL

course, but they do not provide a list of specific course topics. They do promote the idea that the course is important because it introduces the idea of socially and environmentally sustainable IT use.

Topic(s)	Rationale
SECTION 1	
Introduction to ethics and frameworks for ethical analysis and decision-making	Students need basic information about ethical theory and analysis frameworks so they can apply it to IT-specific contexts. They also need to understand how ethics and law are different, but also see how they are related.
Introduction to the legal environment and judicial processes in the US and other jurisdictions (for example, the European Union)	Option for courses where students have not had an introductory business law course. Students should understand the difference between civil and criminal laws and penalties.
SECTION 2	
Privacy	An important topic for individuals and organizations. What data should be collected and how should it be shared and analyzed appropriately.
Freedom of expression and social media	Understand what protections are available for free speech and how this impacts people and organizations involved in social media sites where individuals post information and share opinions.
Intellectual property and cybersecurity	Intellectual property is some of the most valuable organizational assets. They are protected by copyrights, patents, and trademarks, but digitized intellectual property can be more easily stolen online if systems are not secured.
Software engineering	Optional topic for programs where students are being trained to be software developers.
Impact of IT and AI on the global economy and social and environmental sustainability	Broader issues that affect entire industries, labor, the overall global economy, and societal transformation. This topic is directly tied to recommendations from the ACM/AIS IS2020 competency model and the MSIS 2016 report that emphasized socially and environmentally sustainable IT use.
SECTION 3	
Comprehensive review of all of the course topics and how they are interrelated	Previous sections have independently addressed important issues. This concluding section should look at how the topics are interrelated and potential conflicts.

Table 6. Example General ITEL Course Design

Finally, effective ITEL courses should utilize several complementary assessment methods to test students' knowledge and their ability to apply this knowledge to real-world scenarios. Quizzes can be used in each module to assess their factual knowledge about important issues, terminology and laws. In each module, discussion questions and case studies

provide students with scenarios where they can apply their knowledge and look at issues from multiple perspectives focusing on specific real-world issues. They can be asked about what decision they would make, but also asked to describe the rationale behind their decision. Courses can conclude with a comprehensive case where students apply all of their course knowledge and demonstrate that they can recognize ethical problems, legal issues, and potential conflicts between these issues, and then provide solutions to minimize the identified problems. For example, an ITEL course could ask a student to act as a consultant to evaluate a real-world organizational information system or e-commerce site to identify potential ethical or legal problems. Are there any potential unethical uses for the system when it is collecting, storing, and sharing information among its stakeholders? And what general or industry-specific laws (for example, the Health Insurance Portability and Accountability Act (HIPAA) or FERPA) need to be considered.

6. Summary of Findings and Directions for Future Research

This study's findings provide a snapshot for the current state of ITEL class offerings at a sample of US medium sized doctoral universities. It was found that about five out of every eight sampled universities offer at least one ITEL class with some university offering as many as five unique ITEL classes. The courses are most commonly offered at the undergraduate level by computer science departments. Based on an analysis of various university resource, governance, and enrollment profile characteristics, the universities that offer at least one ITEL class tend to have larger total student enrollments, are public universities, and have a majority, or larger proportion, of undergraduate students. The opportunity and ability to offer ITEL courses requires financial and faculty resources. It would be extremely rare for a university to create a new department just so it could offer ITEL courses. Most course descriptions are written in broad terms so they can evolve as specific ethical contexts and laws change over time. A smaller number are more narrowly defined to fit a more specific purpose within a curriculum. It is very rare that specific ethical frameworks or laws are noted in a course description.

This study has some limitations. The sample of universities reviewed is relatively small compared with the large number of US universities and all of the data was collected from university websites and other online sources. These limitations point to some directions for future research. Additional ITEL curriculum review studies could look at samples of larger or smaller US universities to see if the findings from this study are generalizable to these other groups. Studies could also review ITEL course offerings at universities outside of the US. Universities in the European Union (EU) may offer these courses because those countries are very concerned with ethical data management practices and have enacted a number of technology-related laws to protect their citizens and organizations. Another direction for future research could involve digging deeper into the reasons why an individual faculty member or department created an ITEL course. Faculty members that have developed or taught ITEL classes could be interviewed to gain a deeper understanding about what topics they cover, what assessment methods are used, and look at who initiated the course development and the reasons why they felt that the course was needed. These interviews could also focus on the influence that accrediting bodies may, or may not, have on ethics-related curriculum. For example, the Accreditation Board for Engineering and Technology (ABET) (https://www.abet.org/accreditation/) accredits programs in computing technology. And business colleges and schools may be accredited by the Accreditation Council for Business Schools and Programs (ACBSP) (https://acbsp.org/page/accreditation-overview) or the Association to Advance Collegiate Schools of Business (AACSB) (https://www.aacsb.edu/). Each of these accrediting organizations has begun to emphasize social impact and ethics in the past few years and this focus should be expected to continue to grow in the future. The question would be whether this emphasis has impacted curricular decisions in technology-related programs.

7. References

Burns, T. J., Gao, Y., Sherman, C., & Klein, S. (2018). Do the knowledge and skills required by employers of recent graduates of undergraduate information systems programs match the current ACM/AIS information systems curriculum guidelines?. *Information Systems Education Journal*, *16*(5), 56-65.

Carnegie Classification of Institutions of Higher Education (2023), https://carnegieclassifications.acenet.edu/ Accessed August 16, 2023.

Fichman, R. G. (1992, December). Information technology diffusion: A review of empirical research. *Proceedings of the Thirteenth International Conference on Information Systems (ICIS)*, Dallas, 195-206.

Fiesler, C., Garrett, N., & Beard, N. (2020, February). What do we teach when we teach tech ethics? A syllabi analysis. *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289-295.

Föll, P., & Thiesse, F. (2021). Exploring information systems curricula: A text mining approach. *Business & Information Systems Engineering*, *63*(6), 711-732.

Grosz, B. J., Grant, D. G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A., & Waldo, J. (2019). Embedded EthiCS: integrating ethics across CS education. *Communications of the ACM*, *62*(8), 54-61.

Gupta, B., Goul, M., & Dinter, B. (2015). Business intelligence and big data in higher education: Status of a multi-year model curriculum development effort for business school undergraduates, MS graduates, and MBAs. *Communications of the Association for Information Systems*, *36*, 23.

Hirsh, K. J., & Miller, W. (2003). Law school education in the 21st century: Adding information technology instruction to the curriculum. *Wm. & Mary Bill Rts. J.*, *12*, 873.

Leidig, P. M., & Salmela, H. (2022). The ACM/AIS IS2020 competency model for undergraduate programs in information systems: A joint ACM/AIS task force report. *Communications of the Association for Information Systems*, *50*, 25.

Lyytinen, K., Topi, H., & Tang, J. (2021). Information systems curriculum analysis for the MaCuDE project. *Communications of the Association for Information Systems*, *49*, 38.

National Center for Educational Statistics, COLLEGE Navigator, https://nces.ed.gov/collegenavigator/ Accessed August 28, 2023.

Quirk, P. (2008). Curriculum themes: Teaching global cyberlaw. *International Journal of Law and Information Technology*, *16*(3), 297-308.

Reynolds, G. (2019). *Ethics in Information Technology*, 6th Edition, Cengage.

Saltz, J. S., Dewar, N. I., & Heckman, R. (2018, February). Key concepts for a data science ethics curriculum. *Proceedings of the 49th ACM technical symposium on computer science education*, 952-957.

Strader, T., & Bryant, A. (2018). The characteristics of universities offering data analytics programs: An analysis of US regional masters universities. *Journal of the Midwest Association for Information Systems*, 2018(1), 37-48.

Subramanian, R., & White, B. A. (2008) The place of Cyberlaw in MSIS curricula. *Information Systems Education Journal*, 6(29), 3-12.

Topi, H., Karsten, H., Brown, S. A., Alvaro, J., Donnellan, B., Shen, J., ... & Thouin, M. F. (2017). MSIS 2016 global competency model for graduate degree programs in information systems. *Communications of the Association for Information Systems*, *40*(18).

Author Biographies

Troy J. Strader is the Aliber Distinguished Professor of Information Systems and Research Fellow in the Zimpleman College of Business at Drake University. He received his Ph.D. in Business Administration (Information Systems) from the University of Illinois at Urbana Champaign in 1997. Dr. Strader has edited three books and his research has been published in *Decision Support Systems, Electronic Markets, Journal of Information Technology, Journal of the Association for Information Systems, International Journal of Electronic Commerce, Communications of the ACM, Electronic Commerce Research, European Journal of Information Systems* and other academic and practice publications. He is the founding Editor-in-Chief for the Drake Management Review.

J. Royce Fichtner is the Ellis and Nelle Levitt Distinguished Professor of Business Law at the Zimpleman College of Business at Drake University. He received his Bachelor's Degree in Business Administration at the University of Northern Iowa and his Juris Doctorate from Drake University. His major research interests are information technology law, company law, and corporate governance.

This page intentionally left blank

Journal of the Midwest Association for Information Systems

Volume 2024 | Issue 2 Article 3

Date: 07-02-2024

R Code Authorship Attribution using the ASAP Tool

Austin Coursey
Vanderbilt University, austin.c.coursey@vanderbilt.edu

Matthew F. Tennyson
Murray State University, mtennyson@murraystate.edu

Vlad Krotov
Murray State University, vkrotov@murraystate.edu

Abstract

Source code authorship attribution is the task of determining the author of a program. Code authorship attribution has many useful applications, such as plagiarism detection and settling copyright infringement disputes. With the rise in popularity of the R programming language in the Data Science community, the need for source code authorship attribution of R programs has also risen. In this research note, we propose and evaluate the use of a tool called "ASAP: A Source-code Authorship Program" for attributing authorship of R code. We run experiments on two different datasets of R code: a "clean" one (where we are sure of each program's author), and an "unclean" one (with more realistic data, were authorship of some code files is not certain). We find that in both datasets running an experiment using the ASAP tool with the Source Code Author Profile (SCAP) algorithm on R programs attributes authorship successfully. A number of implications for both academics and practitioners are formulated based on these experiments, together with directions for future research in this area.

Keywords: Authorship attribution, SCAP, ASAP, R programming, plagiarism detection, copyright infringement

DOI: 10.17705/3jmwa.000090

1. Introduction

R has become one of the most widely used languages for data analysis in many industries and academic fields (Lander 2014; Schutt and O'Neil, 2014). Academic researchers use R and related tools for collecting, organizing, and analyzing both quantitative and qualitative data (Krotov and Tennyson, 2018). Instructors use R as a tool for equipping students with hands-on knowledge of the research process, statistics, and data wrangling (Fawcett, 2018). Many educational institutions offering Data Science or Business Analytics programs have at least a portion of their curriculum devoted to R. Industry researchers use R for addressing important operational and strategic questions for their organizations in such industries as banking, marketing, pharmaceuticals, public services, etc. (Lander, 2014). In addition, there is a growing practice of building data products and publishing them on the Web with the help of the Shiny add-on to the RStudio Integrated Development Environment (RStudio, 2021a).

Today, there is a vibrant and growing community of R developers coming from all kinds of backgrounds. With R being an "open source" language, there is wide-spread practice of sharing R code freely among the members of the community through such online platforms as GitHub, Stack Overflow, and many others. This code sharing practice creates a problem of proper authorship attribution in the R community. This problem, as explained further in this paper, is driven by two overlapping issues: intellectual property and intellectual honesty. While an open source tool, R can be used by private, for-profit companies to develop proprietary software products. These products constitute intellectual property and are protected by copyright law. However, given the open source nature of R and many R-based projects, the issue of code authorship contribution is usually a problem of intellectual honesty and proper recognition of those who develop the original code. This paper is an attempt towards addressing the problem of proper code authorship attribution in the R community.

This paper explains the problem of authorship attribution of R code given the popularity of R in the Data Science community. The study utilizes a tool (ASAP: A Source-code Authorship Program) for attributing authorship of R code using the Source Code Author Profile (SCAP) algorithm. The primary research question being addressed is this: Can the authorship of R code be properly attributed using methods established for other programming languages? The primary goal is to assist researchers and practitioners in these endeavors by proposing and testing a method and a tool for R code authorship attribution. The tool is tested using large sets of real-world R language project files. These experiments confirm usefulness of the ASAP tool and SCAP algorithm for attributing authorship of R code in real-world scenarios involving students, researchers, and practitioners. Specific ways in which these tools can be used for enforcing intellectual honesty and resolving copyright infringement disputes are proposed. The study also provides directions for future research that can potentially improve accuracy of authorship attribution in the R community.

2. Factors Behind the Growing Popularity of R

Several factors can explain the growing popularity of R among researchers and practitioners. These factors include: the free, open source distribution of R and related tools; built-in functionalities for data analysis; simple and intuitive syntax that facilitates broader use of the language; a comprehensive collection of user-contributed R language extensions that facilitate virtually all known forms of data manipulation, analysis, visualization, and research communication; and presence of vibrant online communities of R developers who share their expertise in R and help others find solutions to R-related problems (R Project, 2021). Each of these factors behind R popularity is discussed in more detail below.

First, unlike some of the well-known proprietary tools for data analysis, such SPSS or SAS, the R language is free and distributed in open-source fashion, in accordance with the terms of the Free Software Foundation's GNU General Public Licence (R Project, 2021). Even some very powerful add-ons for R, such as the RStudio Integrated Development Environment, can be used in an open-source fashion (RStudio, 2021b).

Second, unlike other popular programming languages used in data analysis (e.g., Python, C++, or Java), R was built specifically for data analysis. Just like any programming language, R syntax includes variables, data types, functions for input-output, loops, conditionals, user-defined functions, etc. But on the top of these typical features of a programming

language, R contains built-in facilities for retrieving, cleaning, and storing data; a collection of operators for performing mathematical operations using arrays and matrices; built-in tools for data visualization and statistical analysis; and many other features useful for data analysis (R Project, 2021).

Third, the R language has a simple and intuitive syntax in comparison to other popular and well-established programming languages such as C++ (R Project, 2021). This makes R appealing to users interested in data analysis who, nevertheless, are not professional programmers and may lack a deep understanding of some fundamental aspects of programming, such as the Object-Oriented Paradigm (OOP). While the OOP is implemented in R (Wickham, 2014), most R applications are developed in a script-like, procedural fashion to automate some or all aspects of data retrieval, organization, or analysis (Krotov and Tennyson, 2018). This appeals to statisticians and academic researchers, many of whom can write code but are not professional programmers. Still, R can be used in conjunction with other popular languages, such as Python or C++, to expand its functionality or speed up computations (R Project, 2021).

Fourth, R has a very large collection or R "packages" – user contributed extensions of the R language containing various functionalities for data analysis. For example, the Comprehensive R Archive Network (CRAN) online collection contains almost 17,000 R packages developed by the R user community. These packages contain a wide array of functionalities for processing and analyzing both quantitative and qualitative data – anything from simple data cleaning and manipulation tasks (e.g. "dplyr" package) to advanced machine learning algorithms (e.g. "superml" package). These packages also cover a wide variety of industries and academic fields: from financial analysis (e.g. "XBRL" package) to analysis of literary works performed by academics in the English literature field (e.g. "syuzhet" package). Of special importance are graphical packages (e.g., "ggplot2" package) that contain customizable graphics elements that allow users to create useful and aesthetical data visualizations of virtually all known forms (Lander, 2014). With all these tools or packages available, R can be used to automate and unambiguously describe all steps of a research project: from the time a dataset is loaded from the Web or desktop computer to the time documents and slides are produced to communicate the results. In fact, there is a package called "knitr" that assists in documenting and communicating all the steps and outcomes of a research project (Lander, 2014). These tools for communicating research in R can potentially increase its impact and enhance its reproducibility (Peng, 2011).

Another factor that is probably both a cause and an effect of the growing popularity of R is the presence and constant growth of vibrant online communities of R developers. For example, Stack Overflow, an online community where developers get their programming questions answers and share their solutions and experiences, contains close to half a million answered questions or discussion threads related to R. With these and countless other online communities and resources for R developers, one only has to type "how to do […] in R" to get R source code that does what the user wants in a matter of seconds. Oftentimes, researchers "cut and paste" R code into their own projects without putting much thought into proper attribution of authorship of the code they are using.

While R has all of the advantages outlined above, Python is another popular programming language used in data science. Both R and Python are widely-used and have their advantages and disadvantages. Both are free, relatively easy to learn, and have large reusable libraries. The most fundamental difference between the languages is that Python was designed as a general-purpose programming language, while R was specifically designed for data analytics. R has more packages and specifically more packages that help with data mining and statistical analysis. R can be used to quickly perform specific data analysis tasks and create visualizations, while Python is more suitable to building complete applications (Griffiths, 2022). Both languages are growing in popularity and have their place in the field of data science. In our study, we have chosen to focus on the R langu age.

3. The Problem of R Code Authorship Attribution

This increasing breadth and depth of R usage exacerbates the problem of R code authorship attribution. In this paper we define authorship attribution as the task of deciding who wrote a particular R source file (Zhao and Zobel, 2005). The problem of code authorship attribution, as defined here, is somewhat different and subtler than the problem of detecting plagiarism in R code – a situation where somebody does a "copy and paste" of R code written by somebody else and presents it as his or her own solution. The problem of code plagiarism has been addressed in the field of Computer Science with the help of such tools as JPlag (Prechelt et al., 2002) and MOSS. Moreover, the traditional plagiarism detection tools, such as Turnitin or SNITCH, although not designed specifically for code plagiarism detection, can detect "copy paste" code

as well (Niezgoda and Way, 2006). The problem of R code authorship contribution arises when someone reuses R code by modifying it for his or her own research project or proprietary data product. In this case, the code is not plagiarized (at least, not in a blatant fashion). But the problem of determining who is the author of this modified code still remains.

This problem of deciding on the true author of an R code is important for both academics and practitioners. While R development is often guided by the open-source practices (RStudio, 2021b), the problems of fairness to the original author of code and intellectual honesty still remain. Just because the code is open source, this does not mean that someone can take it and present it as his or her own. For example, most of the licenses under Creative Commons, an organization responsible for setting licensing level in the open-source community, require proper authorship attribution when open-source code is used (Creative Commons, 2021).

It is interesting that nowadays many academic journals encourage or even require the submission of supplementary resources (e.g., data sets, software code, etc.) together with the text of an article itself. But it is usually only the article that is checked for plagiarism and not the source code that was used to conduct a study that led to the article. With the availability of such R tools as "knitr" (a package that allows to "mesh" together R code, analysis output, and author text explaining the analysis and interpreting the findings), the boundary between an article and the code used to generate the article may become quite blurry.

Another aspect of the problem facing academics in relation to R code authorship involves students submitting a modified version of somebody else's R code for course credit. Opinions regarding the appropriateness of code reuse for classroom projects may differ among educators. Still, most academics, being exposed to the topics of plagiarism and intellectual honesty in their doctoral programs and via policies of the professional organizations that they belong to, would still prefer students to acknowledge that they are relying on somebody else's work for their classroom projects.

When it comes to proprietary R products and solutions, then the problem of R code authorship transcends the realm of ethics and becomes a legal issue. As most developers know, software constitutes intellectual property and can be copyrighted (Bainbridge, 1999). Intellectual property in the form of software code constitutes a very strategic intangible asset for modern companies (Lev, 2000). Improper use of copyrighted intellectual property, be it text, R code, or even data that that the text or R code is related to, constitutes copyright infringement – something that can lead to a lengthy and costly litigation in a court of law (Dreyer and Stockton, 2013).

Given the growing popularity of R in industry and academic research projects, both academics and practitioners need to become well-equipped for enforcing the highest standards of intellectual honesty and protecting their intellectual assets. The main goal of this paper is to assist researchers and practitioners in these endeavors by proposing and testing a method and a tool for R code authorship contribution.

4. Additional Literature Review

Determining authorship of natural language text is a problem that has been studied for years across many different applications. Stamatatos (2008) provided brief history of authorship attribution of natural language documents with a focus on modern computational methods. In addition to a review of the history of authorship attribution, an overview is given of stylometric features commonly used in natural language processing and an overview of approaches used, such as profile-based, probabilistic, compression, similarity-based, and hybrid models.

In the past year alone, several studies of authorship attribution of natural language documents have been published. Makhmutova et al. (2023) evaluated the importance of attention scores using character n-grams specifically within news articles written in English and Russian. Uchendu et al. (2023) provided a review of not only authorship attribution methods, but also authorship obfuscation methods, where "authorship obfuscation" refers to the task of modifying a document to hide its true authorship. Alqahtani and Dohler (2023) looked at the problem of authorship attribution specifically of Arabic documents, which is uniquely challenging due to the complexity of the Arabic-language morphology. Sebastián et al. (2023) presented a method for identifying the owner of a domain or website. Their method was shown to be much more accurate than the traditional approach of using the "WHOIS" protocol for looking up a website's registered owner. These recent publications provide just a sampling of the ongoing research and innovation within the field of natural language authorship

attribution.

Determining the authorship of programs and the related problem of detecting plagiarism of source code is a much newer area of study. Kalgutkar et al. (2019) provided a summary of code authorship methods and challenges. They specifically frame the problem around the challenge of identifying the source of malware as their main motivation. A plethora of studies have been published in regards to various aspects of source code authorship attribution. Methods have been proposed to attribute authors of programs written in traditional programming languages like Java and C++ and studies have measured the effectiveness of those methods (Frantzeskou et al., 2007; Burrows and Tahaghoghi, 2007; Ding and Samadzadeh, 2004; Tennyson, 2013). Tools have been created to detect plagiarism of source code (Prechelt et al., 2002), especially within the context of programming assignments in the classroom. Studies have looked at source code authorship attribution from a practical software engineering standpoint (Bogomolov et al., 2021). Petrik and Chuda (2021) studied how a programmer's style changes over time and how that affects source code authorship attribution. Dauber et al. (2018) looked at how to attribute authorship of small, incomplete code fragments. Gonzalez et al. (2018) proposed an approach to perform authorship attribution specifically on Android apps. Hendrikse (2017) investigated the possibility of attributing authorship of executable machine code, when the source code isn't available, with surprising levels of accuracy.

So, authorship attribution is a problem that has been studied in many different ways and in many different applications for many different languages. However, to our knowledge, it has not been studied specifically in R code. That being said, studies have been performed in regard to intellectual property of open source software, which is a related area of study. Santos et al. (2011), for example, executed a longitudinal study to evaluate the impact of changes in intellectual property policy of open source projects. Additionally, studies requiring the analysis of R code from the CRAN archive have been published, but not specifically related to authorship attribution. For example, Atchison et al. (2018) utilized R code from the CRAN archive to analyze the topic space of scientific computing.

5. The ASAP Tool

The tool proposed for R source code authorship attribution is the ASAP tool (Tennyson, 2019). It was chosen because of its ease of use, variety of experimental settings, and proven performance. To install the ASAP tool, all one must do is download the ASAP folder to any location on the target computer. The ASAP folder can be obtained and downloaded from the following GitHub repository: https://github.com/ASAP-Project/ASAP.

The ASAP tool requires a Java runtime engine and Perl. If the user does not have Java or Perl installed, those can be obtained for free online. Additional instructions are provided in the tool's documentation. Once the user has Java and Perl, all that must be done to run the tool is execute the ASAP.jar file.

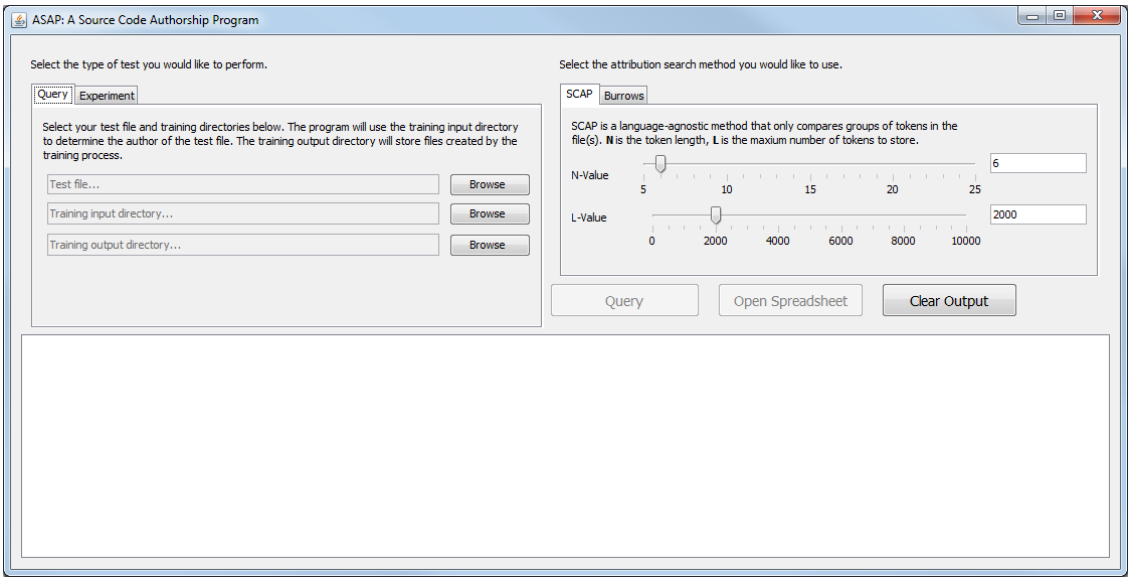


Figure 1. ASAP GUI

Once the ASAP.jar file is executed, a GUI will appear (see Figure 1). This GUI contains the options for the query or experiment the user can run. First, the user can choose between running a query or an experiment. A query will test for the author of a single unknown file. If a query is selected, the user must select their test file, their input directory, a directory containing folders from the possible authors with their known source code inside, and the output directory. An experiment will attribute the author to many documents at once. If an experiment is selected, one of three different types of experiments can be chosen. The first is a default-split experiment. This "queries all the files in the test directory, using the files in the training directory to attribute authorship." (Tennyson, 2019) The next is a k-fold experiment. This splits the dataset into k number of folds. One of these folds is used at a time to query while the others are used to attribute authorship. The final type of experiment supported by the ASAP tool is a leave-one-out experiment. This tests every file, one at a time, against every other file in the test directory. For a more detailed description of the types of test the ASAP tool offers, the reader is encouraged to read Tennyson (2019).

After a test type is determined, the user must then determine the method of authorship attribution they would like to use. The ASAP tool provides two options for this: the SCAP method (Frantzeskou et al., 2007) and the Burrows method (Burrows and Tahaghoghi, 2007). If the SCAP method is chosen, the user has the option to change the token length and maximum number of tokens to store. A more in-depth description of the SCAP method is provided further in the paper.

The Burrows method is not used in this study. The main reason for excluding the Burrows method is that it depends on the programming languages being analyzed: its search utilizes features of the specific programming language, such as keywords, identifiers, operators, etc. In order to apply the Burrows method to the R programming language, features would have to be selected for the R language, which is outside the scope of this study. Due to its current lack of support for the R programming language, an explanation of the Burrows method and its settings supported by the ASAP tool are omitted.

Author	NumFiles	NumCorrect	Percentage
AaronRobotham	24	20	83.3%
AbdulMajedRajaRS	5	4	80.0%
AdamLund	12	11	91.7%
AdamRothman	20	20	100.0%
AdelinoFerreiradaSilva	25	23	92.0%
AlbertoKroneMartins	17	16	94.1%
AlboukadelKassambara	9	9	100.0%
AlexCannon	53	51	96.2%
AlexZvoleff	14	14	100.0%
AlexisDinno	9	9	100.0%

File	AaronRobotham	JulMajedRaj	AdamLund	damRothma	noFerreirad	rtoKroneMa
addhead.R	47	5	12	15	15	18
car2sph.R	101	12	23	33	35	29
deg2dms.R	97	18	32	36	42	31
deg2hms.R	92	12	21	32	31	27
dms2deg.R	117	18	50	68	53	43
genparam.R	47	10	40	26	36	57
hms2deg.R	107	17	30	54	49	32
IAUID.R	48	9	12	16	12	13

Figure 2. ASAP Excel spreadsheet output samples. Attribution totals (top). Each file's attribution for a specific author (bottom).

Once the test type and method of authorship attribution are determined, the user can simply click the "Query" or "Experiment" button and the Java GUI will run the necessary Perl commands to run the test. Once the experiment is finished, an Excel spreadsheet will be generated with the results (see Figure 2), and the user can click the "Open Spreadsheet" button to view it. Additionally, if the user wishes to bypass the GUI, such as in the case of running an experiment with an L-value larger than 10,000, then the user can run the Perl commands themselves from the command line.

6. The SCAP Method

The method of source code authorship attribution chosen for this project was the SCAP method. It was chosen over the other method incorporated by the ASAP tool, the Burrows method for reasons mentioned above.

The SCAP method, unlike the Burrows method, is not dependent on any set programming language features. Thus, it can be used with any programming language. It also has proven high performance in the C++ and Java languages (Frantzeskou et al., 2007). All of these factors contributed to our choosing of this method for testing authorship attribution in R. A brief overview of how the SCAP method works is provided below.

The SCAP method of source code authorship attribution creates a unique profile for each author. The program whose author is unknown is compared against each author's Source Code Author Profile (SCAP) using a similarity measure known as the Simplified Profile Intersection (SPI). The SCAP that is the most similar to the program is attributed authorship of that program.

An author's profile is created by concatenating together all of the available programs known to be written by that author. Instead of representing this profile as plain text, it is represented as a table of byte-level n-grams. Since these n-grams are extracted at the byte level, the SCAP method is not dependent on any programming language. All the characters in all the files by an author are converted into n-grams. Only the L most frequent n-grams are kept in the table, and this table becomes the author's profile.

The SPI is "the number of n-grams an author profile and a program have in common" (Tennyson, 2019). For a program being tested, the author that has the highest SPI with it is attributed authorship.

7. Methodology

To test the effectiveness of the SCAP method of authorship attribution with the R programming language using the ASAP tool, two experiments were run. First, an experiment on clean data was run. This featured a dataset that we knew contained R programs written by specific individual authors. Second, an experiment was run on unclean data. This was done to replicate what real-world data might look like. It contained any R programs we could find, even ones with multiple or unclear authors. All of the experiments were executed on a machine with an Intel Core i7 CPU running at 1.8 GHz with 4 cores and 16 GB of RAM. A more detailed description of how the data for these experiments were gathered and the ASAP settings chosen is provided below.

7.1. Gathering and Cleaning the Data

To gather the data for the clean experiment, the CRAN R Project Archive (CRAN Archive, 2021), one of the most popular repositories of R code, was chosen as the source of the R programs. It was chosen primarily because of its very large repository of R programs and its documentation of the authorship of its packages.

From this repository, one unique author with more than 2 packages published was chosen at a time. Only authors with more than 2 packages published were chosen to limit the number of authors and prevent authors with too few files from being included in the dataset. If an author had a similar name to another author, the author with fewer packages was excluded. If a package had multiple authors, that package was not included in the dataset. These steps were done to ensure that each author profile would represent a distinctive author.

Once an author was selected, its package names were manually reviewed to see if any had the same base name. If any did, they were excluded from the dataset. For example, "package1" and "package2" both have the base name, "package", so they would both be excluded. If more than two usable packages still existed by that author, then those packages were downloaded from the CRAN Archive into a folder named after that author. In the case that an author had more than 2 usable packages and the download failed to acquire 1 or more of them, the author was kept in the dataset as there is another step further on that handles authors with too few files. A total of 855 packages from 272 authors were downloaded.

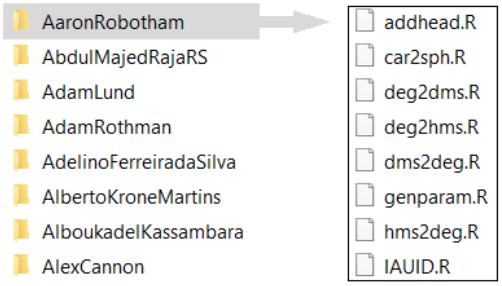


Figure 3. File structure. Author folders all in the same folder. Inside each author folder are the files that belong to that author.

After all the available authors and their usable packages were downloaded, the author folders contained compressed packages in a tar.gz format. Those were programmatically decompressed and unpackaged. The packages by an author, now in a normal folder format, contained many files pertaining to testing and documentation of those packages. The actual R files for that package were in a folder called "R". Anything that was not in that folder was programmatically deleted. The files inside that folder were brought out of that folder and that folder was deleted. The R files were then brought out of their respective packages into the author folder, and the now-empty packages were deleted. This led to the author folders containing only the R programs written by that author, the file structure needed for the ASAP tool (see Figure 3).

To ensure that none of the programs were duplicates of each other in the entire dataset, we programmatically checked the similarity of the text in the files. If any of the files were more than 95% similar to another, they were considered duplicate and both removed from the dataset. In the case that one file was more than 95% similar to another but the second file was less than 95% similar, they were still both excluded. In total, 33 files were found to be duplicates and were excluded. It should be noted that this is much different than the process of attributing authorship. This step simply compared the text of the programs while source code authorship attribution focuses on more specific details such as stylistic features of an author.

With the duplicate files gone, any author with fewer than four R programs was programmatically deleted. This was to ensure that each author had enough data to build a profile that represented them as accurately as possible. In total, there were 6,197 R files representing 225 authors for the clean dataset.

To gather the data for the unclean experiment, we simply programmatically downloaded as many authors and their respective packages from the CRAN Archive as possible. If a package was authored by more than one person, those people together were treated as a unique author. If an author appeared more than once with a slightly different name, we treated both of those appearances as different authors.

Not much cleaning was done on the data in comparison with the data used for the clean experiment. The packages were decompressed and put into the same file structure as the clean data using the same programmatic approaches. Any author with fewer than four files was removed from the dataset to ensure they would have an accurate profile. This led to a total of 5,674 authors and 96,074 R files for the unclean dataset.

7.2. Running the Clean Experiment

Three different experiments were run on the clean dataset. The only variable changed throughout these was the L-value, the length of the author profile, to test how that length would impact the results.

The first experiment was run on the dataset of 225 authors and 6,197 R files using a leave-one-out experiment with the SCAP method incorporated in the ASAP tool. A leave-one-out experiment was chosen to ensure that every program got individually compared against every author profile to produce the most accurate results; it maximizes sampling while avoiding overfitting.

An n-gram length of 6 bytes and an L-value of 690,000 were chosen. That L-value was selected because previous research showed that the higher the L-value, the better the attribution accuracy (Tennyson and Mitropoulos 2014). The L-value was almost 1,000 over the number of bytes the largest author folder had, 689,003. This ensured that every n-gram from every author would be represented. The ASAP GUI has a max L-value of 10,000, so to use the L-value of 690,000, we ran the command generated by the ASAP tool from the command line with the L-value we chose instead of the default 2,000.

The second and third experiments were run using the same leave-one-out experiment using the SCAP method. The ASAP GUI was used in these experiments with an n-gram length of 6 bytes and an L-value of 10,000 and 5,000 respectively. These were done to see how a smaller L-value might impact the accuracy of the authorship attribution.

7.3. Running the Unclean Experiment

Only one experiment was run on the unclean data. After attempting to run a leave-one-out experiment with an L-value of 10,000, we realized that including all 5,674 authors and 96,074 files was not computationally feasible for us. The reason it was not feasible is explained in the next paragraph.

A leave-one-out experiment removes one file for testing and performs training using all of the other files. This is done for every file in the data set, thus maximizing the number of tests as well as the size of the training set for each test. So, in this case, the top 10,000 n-grams of 96,074 files would be compared to the top 10,000 n-grams of 96,073 other files. This leads to a total of 9,230,117,402 file comparisons. If 100 comparisons could be done a second, this would take over 1,068 days to complete. (9,230,117,402 / 100 comparisons / 60 seconds / 60 minutes / 24 hours = 1,068.3). As a note, 100 comparisons per second is not necessarily representative of the speed of an experiment. The speed will vary based on the processing power of the computer running it; we observed less than 100 comparisons per second. To make the data possible to experiment on, we excluded files and authors from the dataset.

The steps for fairly shrinking the clean dataset are as follows and were done programmatically: remove any author with fewer than ten files, randomly remove authors until only 300 are left, randomly remove files from each of those authors until each author is left with only 10 files. This left the unclean dataset with 300 authors, 3,000 R programs, and 8,997,000 file comparisons - something much more realistic to run an experiment on.

After this, a leave-one-out experiment with an L-value of 10,000 and n-gram length of 6 bytes was run. These are the same parameters as the second experiment run on the clean dataset.

8. Results

The results of running a leave-one-out experiment using the SCAP method with different L-values on the clean dataset, where we knew each file belonged to its respective author, can be seen in Table 1. As is shown, the L-value 690,000 resulted in 83.59% of files being correctly attributed to their author. When the profile length was 10,000 n-grams, 92.29% of files were correctly attributed. Finally, with an author profile length of 5,000 n-grams, 91.22% of the files were attributed correctly.

Table 1. Results of using the SCAP method with different L-values on clean data.

N-Gram Length (bytes)	L-Value (n-grams)	Number of Files	Number Correctly Attributed	Percentage Correct
6	**690,000**	**6,197**	**5,180**	**83.59%**
6	**10,000**	**6,197**	**5,719**	**92.29%**
6	**5,000**	**6,197**	**5,653**	**91.22%**

For the experiment with an L-value of 10,000, a 5-number-summary of the percentage of each author's files correctly attributed is shown in Table 2. An accompanying box plot can be seen in Figure 4. The impact of the number of files each

author has in the dataset on the percentage of their programs correctly attributed can be seen in Figure 5. The number of files has been natural-log-transformed to better show this relationship.

Table 2. Summary of author files correctly attributed using L=10,000 on clean data.

5-Number Summary	
Min	**0%**
Q1	**82.35%**
Median	**93.33%**
Q3	**100.00%**
Max	**100.00%**

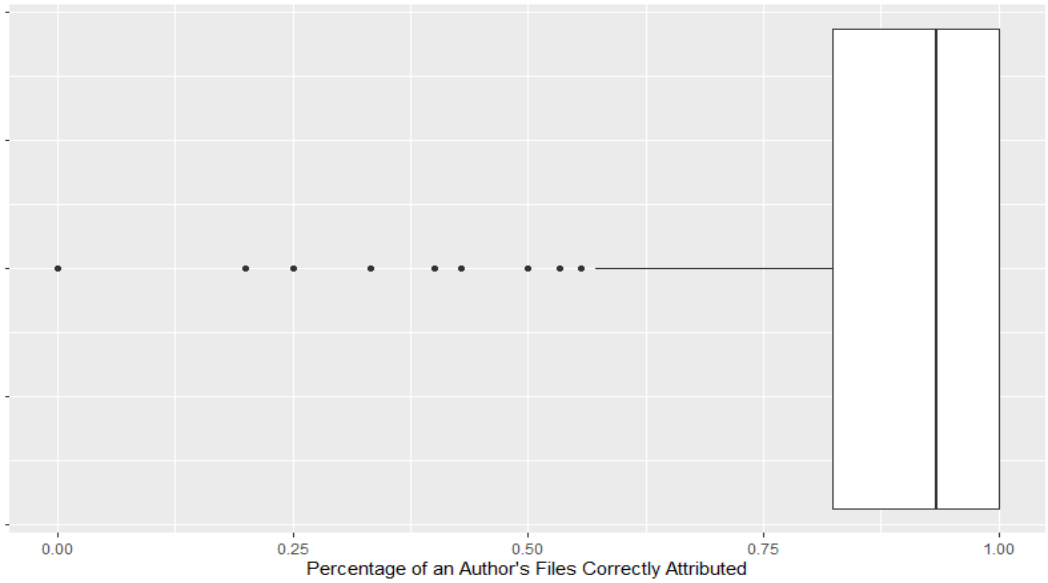


Figure 4. Box plot displaying the 5-number summary from Table 2.

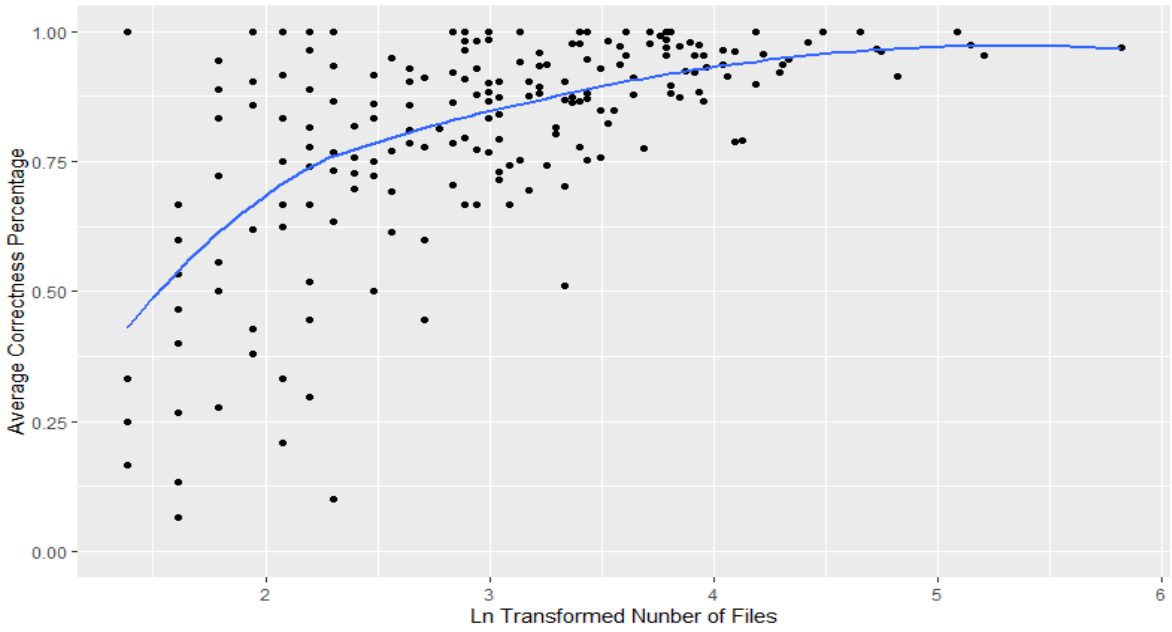


Figure 5. The percentage of each of the 10,000 n-gram length authors' files correctly attributed vs the ln-

transformed number of files that author has.

The results of running a leave-one-out experiment using the SCAP method with an L-value of 10,000 on the unclean dataset, where we could not be sure each file belonged to its respective author, can be seen in Table 3. As is shown, 2,682 out of 3,000 R programs were attributed to their correct author. A 5-number-summary and accompanying boxplot are also shown for this experiment in Table 4 and Figure 6, respectively.

Table 3. Results of using the SCAP method with an L-value of 10,000 on unclean data.

N-Gram Length (bytes)	L-Value (n-grams)	Number of Files	Number Correctly Attributed	Percentage Correct
6	10,000	3,000	2,682	89.4%

Table 4. Summary of author files correctly attributed using L=10,000 on unclean data.

5-Number Summary	
Min	40%
Q1	80.00%
Median	90.00%
Q3	100.00%
Max	100.00%

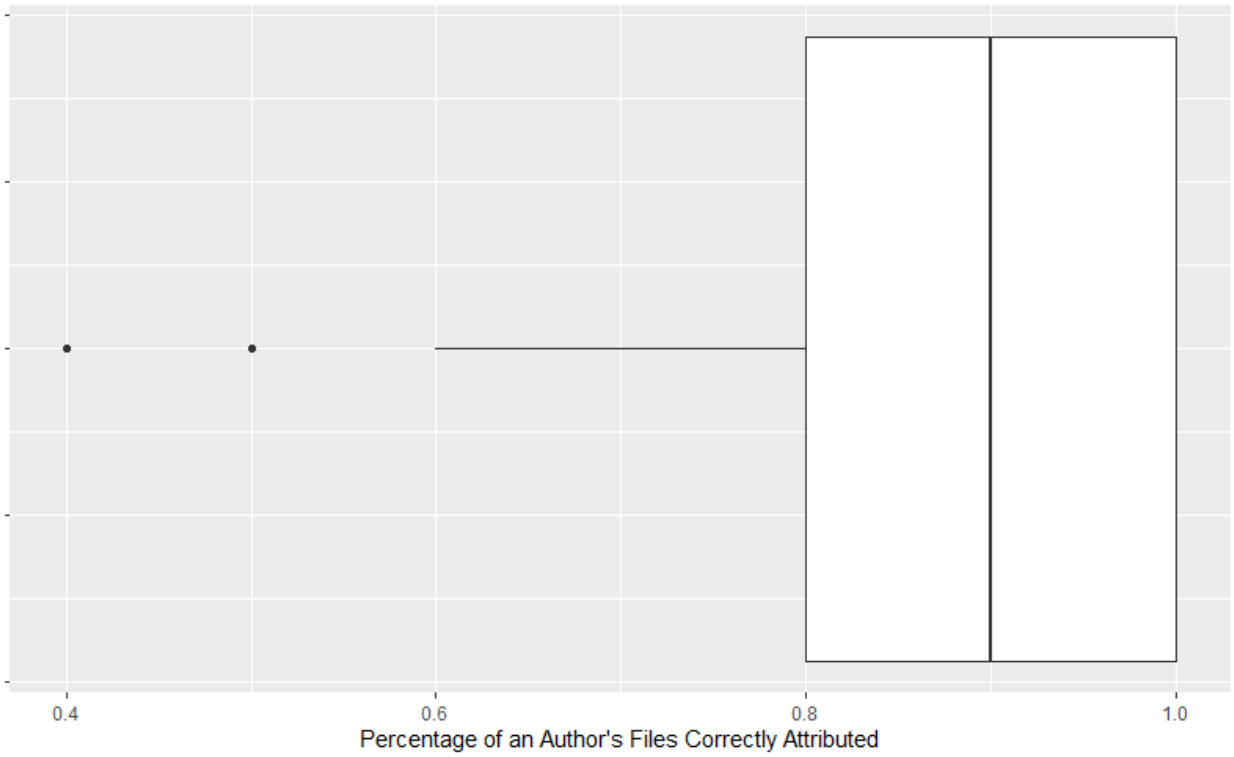


Figure 6. Box plot displaying the 5-number summary from Table 4.

9. Analysis of Results

The leave-one-out SCAP method experiments run on the "clean" dataset correctly attributed different percentages of the files based on how large the L-value was. As can be seen, the experiment with an L-value of 10,000 performed better than a virtually infinite L-value and a smaller L-value. This was unforeseen by us because previous research (Tennyson and Mitropoulos, 2014) implied that the larger the L-value, the more accurate the results.

In that experiment with an L-value of 10,000, 92.29% of the R programs were attributed to their correct author. This is comparable to results from previous research using the SCAP method on Java and C++ programs (Frantzeskou et al., 2007; Tennyson and Mitropoulos, 2014).

Despite the increase in profile length seeming to have a negative impact on the results of the experiment, increasing the number of files per author seems to improve the accuracy of that author's attribution. This is quite intuitive: the more samples of code you have for someone, the more accurate of a profile you can build for them.

The results of the leave-one-out SCAP experiment with an L-value of 10,000 on the "unclean" dataset showed 89.4% accuracy. As was expected, this is a lower accuracy than the one achieved with the "clean" dataset. In the "unclean" dataset, which more closely follows real-world data, one cannot be sure that any given R program was written by the same sole author as the rest of the programs in the author profile. This leads to less accurate profiles. As an example, if an author profile consists of nine R programs written by one author and one R program written by another, the author profile will consist mostly of n-grams from the author with nine R programs. If the programming style of the author with only 1 R program in that package is different than the other author, then that R program will likely not be correctly attributed to that package.

Although the experiment with the "unclean" data did not perform as well as the "clean" dataset, it still performed relatively well. Nearly 90% of the files were attributed to their correct author.

10. Implications

In an academic setting, the need for determining the author of an R program is clear. First, it is needed to ensure that students did not plagiarize or collaborate with one another on an assignment. Second, it could be needed by academic journals to maintain integrity and ensure that everyone is getting the credit they deserve. The need for attributing authorship of R code also exists outside of an academic setting. It is easy to imagine many situations where companies producing products that use R would want to ensure their products are wholly written by their employees for both legal and ethical reasons.

The way to determine if a program was written by a student in one's class, a researcher submitting to one's journal, or an employee at one's place of work is not as simple as just checking for blatant plagiarism. Comparing files to see what percentage of one was copied from the other is straightforward, and there are many tools available to check for this. The real problem arises when someone copies another's code but modifies it. The SCAP method of authorship attribution integrated in the ASAP tool can be used to address such situations in academic and industry contexts.

For example, one can imagine a scenario where an instructor keeps all of the assignments submitted by students so far, both individual and group assignments. The instructor could then run each file submitted for the current assignment through the ASAP tool, using the previous assignments to build an author profile for each student. If the students have worked together before, or one contributed much more to their current assignment they were not supposed to collaborate on, one of their R programs will likely not be attributed to the student that submitted it. The instructor, seeing this, could then delve deeper into the code to make a judgement.

In a similar fashion, a journal editor can build a collection of profiles of R code authors using one or more popular R code repositories such as CRAN (used in this study), GitHub, or Stack Overflow. Given that the journal requires submitting R

code used for data retrieval, processing, and analysis – the submitted code can be compared to existing profiles of known R code contributors to see whether it has a high match with one or more authors (other than the ones who authored the paper). If it does closely match the code of other R code authors, the editor can make sure that the authorship of the code used in the study is properly attributed to avoid potential issues with intellectual honesty or even copyright infringement.

In the workplace, an employer could utilize the ASAP tool with the SCAP method to build profiles for its employees or company as a whole. This could include a routine check on the codebase against a collection of open-source projects to ensure that the company is not, knowingly or unknowingly, using the code of someone else. It could also potentially be used as evidence in cases of copyright infringement. Since the SCAP method builds profiles based on the programming styles of the authors, a profile could be built for the company as a whole, which would hypothetically reflect their coding standards and the style of the employees as a collective. If a suspect R program is more similar to Company A's profile than Company B's profile, then it would be more likely written by Company A.

11. Future Work

Further work could be done to improve upon the knowledge of the effectiveness of authorship attribution with the R programming language and authorship attribution as a whole. One such example is determining what the optimal L-value for the SCAP method is. As is shown in the results of the running the experiment on the "clean" data, the percentage of correct attribution is dependent on the number of n-grams included in the author programs, the L-value. While the L-value of 10,000 yielded good results for our experiment, there is likely a value that is better, and discovering that could improve the accuracy of authorship attribution in R. Another possible task for future research is determining a feature set to be used to apply the Burrows method of authorship attribution to the R programming language. Generative AI is increasingly being used to develop software, including R code. It would be interesting to study how generative AI affects the problem of authorship attribution. A final example could be running more experiments on real-world data with differing files per author. To shrink our "unclean" dataset without bias, we kept the number of files per author consistent, but it is possible that the results of the experiment would have been slightly different if the number of files per author had remained variable like in the "clean" dataset. In a real-world application, there are scenarios where the number of files per author will be both variable and consistent, so determining the effect of the variability in file number could improve the accuracy of R authorship attribution.

12. Conclusion

With the continual rising popularity of R for students, academic researchers, and industry Data Scientists - the need for determining the author of an R program is also increasing. More and more people are needing to write code in R for their personal, work, or academic purposes. Along with this, the amount of R resources available with a simple online search are increasing by the day. The practice of "copying and pasting" someone else's code into one's own research project and modifying it for the project's unique purposes is commonplace. This gives rise to the ethical and legal problem of deciding who is the actual author of the code.

We recommend the use of the ASAP tool based on the SCAP for one's R authorship attribution needs. It is a program that is simple to install, comes with an intuitive GUI, and implements the SCAP method that has shown successful performance attributing authorship to R code. As is shown in previous sections of the paper, the SCAP method using a leave-one-out experiment and an L-value of 10,000 was able to correctly attribute authorship of R programs 92.29% of the time. Even when files were just thrown into the experiment with no confidence that each author profile consisted of programs written solely by that author, it was able to correctly attribute authorship 89.4% of the time. Thus, we deem this method of R code authorship contribution to be useful for many real-world applications.

13. References

Alqahtani, F. and Dohler, M. (2023). Survey of Authorship Identification Tasks on Arabic Texts, *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(4), 1-24.

Atchison, A., Anderson, H., Berardi, C., Best, N., Firmani, C., German, R., and Linstead, E. (2018). A topic analysis of the R programming language, *Proceedings of the 40th International Conference on Software Engineering: Companion Proceedings*, 183-184.

Bainbridge, D. I. (1999). *Software Copyright Law*. West Sussex, UK: Bloomsbury Professional.

Bogomolov, E., Kovalenko, V., Rebryk, Y., Bacchelli, A., and Brykson, T. (2021). Authorship attribution of source code: a language-agnostic approach and applicability in software engineering, *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 932-944.

Burrows, S. and Tahaghoghi, S.M.M. (2007). Source code authorship attribution using n-grams, *Proceedings of the 12th Australasian Document Computing Symposium*, 32-39.

CRAN (2020). Retrieved from https://cran.r-project.org/web/packages/

CRAN Archive (2021). Retrieved from https://cran.r-project.org/src/contrib/Archive/

Creative Commons (2021). Three "Layers" of Licenses. Retrieved from https://creativecommons.org/licenses/

Dauber, E., Caliskan, A., Harang, R., and Greenstadt, R. (2018). Git blame who? Stylistic authorship attribution of small, incomplete source code fragments, *Proceedings of the 40th International Conference on Software Engineering: Companion Proceedings*, 356-357.

Ding, H. and Samadzadeh, M. (2004). Extraction of java program fingerprints for software authorship identification, *The Journal of Systems and Software, 72*, 49-57.

Dryer, A.J. and Stockton, J. (2013). Internet "Data Scraping": A Primer for Counseling Clients. New York Law Journal. Retrieved from https://www.law.com/newyorklawjournal/almID/1202610687621

Fawcett, L. (2018). Using Interactive Shiny Applications to Facilitate Research-Informed Learning and Teaching. *Journal of Statistics Education, 26*(1), 2-16.

Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C., and Howald, B. (2007). Identifying Authorship by Byte-Level N-Grams: The Source Code Author Profile (SCAP) Method. *International Journal of Digital Evidence, 6*(1).

Gonzalez, H., Stakhanova, N., and Ghorbani, A.A. (2018). Authorship Attribution of Android Apps, *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy*, 277-286.

Griffiths, T. (2022). Python vs. R: What's the Difference? Retrieved from https://www.indeed.com/career-advice/career-development/r-vs-python

Hendrikse, S. (2017). *The Effect of Code Obfuscation on Authorship Attribution of Binary Computer Files*, Nova Southeastern University, Fort Lauderdale-Davie, FL.

Kalgutkar, V., Kaur, R., Gonzalez, H., Stakhanova, N., and Matyukhina, A. (2019). Code Authorship Attribution: Methods and Challenges, *ACM Computing Surveys, 52*(1), 1-36.

Krotov, V. and Tennyson, M. (2018). Research Note: Scraping Financial Data from the Web Using the R Language. *Journal of Emerging Technologies in Accounting, 15*(1), 169-181.

Lander, J. P. (2014). *R for Everyone: Advanced Analytics and Graphics*. Boston, MA: Addison-Wesley.

Lev, B. (2000). *Intangibles: Management, measurement, and reporting*. Brookings institution press.

Makhmutova, L., Ross, R., and Salton, G. (2023). Impact of Character n-grams Attention Scores for English and Russian News Articles Authorship Attribution. *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, 939-941.

Niezgoda, S. and Way, T. P. (2006). SNITCH: a software tool for detecting cut and paste plagiarism, *ACM SIGCSE Bulletin, 38*(1), 51-55.

Peng, R. D. (2011). Reproducible research in computational science, *Science, 334*(6060), 1226-1227.

Petrik, J. and Chuda, D. (2021). The effect of time drift in source code authorship attribution: Time drifting in source code – stylochronometry, *Proceedings of the 22nd International Conference on Computer Systems and Technologies*, 87-92.

Prechelt, L., Malpohl, G., and Philippsen, M. (2002). Finding plagiarisms among a set of programs with JPlag. J. *UCS, 8*(11), 1016.

R Project (2021). What is R? Retrieved from https://www.r-project.org/about.html

RStudio (2021a). Shiny from RStudio. Retrieved from: https://shiny.rstudio.com

RStudio (2021b). Take control of your R code. Retrieved from https://www.rstudio.com/products/RStudio/

Santos, C.D., Cavalca, M.B., Kon, F., Singer, J., Ritter, V., Regina, D., and Tsujimoto, T. (2011). Intellectual property policy and attractiveness: a longitudinal study of free and open-source software projects, *Proceedings of the ACM 2011 conference on computer supported cooperative work*, 705-708.

Schutt, R. and O'Neil, C. (2013). *Doing data science: Straight talk from the frontline*. Sebastopol, CA: O'Reilly Media, Inc.

Sebastián, S., Diugan, R.G., Caballero, J., Sanchez-Rola, I., and Bilge, L. (2023). Domain and Website Attribution beyond WHOIS, *Proceedings of the 39th Annual Computer Security Applications Conference*, 124-137.

Stack Overflow (2021). Questions tagged [r]. Retrieved from https://stackoverflow.com/questions/tagged/r

Stamatatos, E. (2008). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538-556.

Tennyson, M. (2013). *Authorship Attribution of Source Code*, Nova Southeastern University, Fort Lauderdale-Davie, FL.

Tennyson, M. (2019). ASAP: A Source Code Authorship Program. *International Journal on Software Tools for Technology Transfer*.

Tennyson, M. and Mitropoulos, F. (2014). Choosing a profile length in the SCAP method of source code authorship attribution. *IEEE SOUTHEASTCON 2014*. Lexington, KY, USA, 13-16 March 2014. IEEE.

Uchendu, A., Le, T., and Lee, D. (2023). Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective, *ACM SIGKDD Explorations Newsletter*, *25*(1), 1–18.

Wickham, H. (2014). *Advanced R*. Boca Raton, FL: CRC Press

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Asia Information Retrieval Symposium*, 174-189.

Author Biographies

Austin Coursey is a Computer Science Ph.D. student in the Institute for Software Integrated Systems at Vanderbilt University. He earned his B.S. in Computer Science and Mathematics from Murray State University in 2022. His research applies and develops deep learning techniques to solve challenging problems in complex systems. His recent work has addressed robust unmanned aerial vehicle control using reinforcement learning, continual reinforcement learning, anomaly detection for building energy consumption and freeway traffic incidents, and data-driven prognostics for aircraft engines and hard disk drives. He authored this work during his time as an undergraduate student at Murray State University.

Dr. Matthew Tennyson is an Associate Professor of Computer Science at the Department of Computer Science and Information Systems, Arthur J. Bauernfeind College of Business, Murray State University. Matthew earned his B.S. in Computer Engineering from Rose-Hulman in 1999. He worked at Caterpillar as an engineer, developing embedded systems for various types of earth-moving machinery. He earned his M.S. in Computer Science from Bradley University. He earned his Ph.D. at Nova Southeastern University in 2013. Matthew's teaching and research interests include software engineering, programming practice and theory, and computer science education.

Dr. Vlad Krotov is a Professor of Information Systems and Cybersecurity Management at the Department of Computer Science and Information Systems, Arthur J. Bauernfeind College of Business, Murray State University. Dr. Vlad Krotov received his PhD in Management Information Systems from the Department of Decision and Information Sciences, University of Houston (USA). His teaching, research and consulting work is devoted to helping managers and organizations to use Information and Communication Technologies for analyzing organizational data in a way that enhances organizational performance. His quantitative and qualitative research has appeared in a number of academic and practitioner-oriented journals and conferences, such as: CIO Magazine, Journal of Theoretical and Applied E-Commerce, Communications of the Association of Information Systems, Business Horizons, Blackwell Encyclopedia of Management, America's Conference on Information Systems (AMCIS), Hawaii International Conference on System Sciences (HICSS), International Conference on Mobile Business (ICMB). His research was recognized by the 2016 Outstanding Researcher award and 2017 Emerging Scholar award at Murray State University.