12-31-2024

Interactive visualization workflows for mitigating analytical uncertainty

Kaustav Bhattacharjee
New Jersey Institute of Technology, kaustavbhatt94@gmail.com

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

INTERACTIVE VISUALIZATION WORKFLOWS FOR MITIGATING ANALYTICAL UNCERTAINTY

by
Kaustav Bhattacharjee

This dissertation takes a process-centric and stakeholder-first perspective for handling analytical uncertainty: the form of uncertainty that confronts data analysts' insight-generation processes in high-consequence decision-making scenarios. The cost of an incorrect decision when data is used for movie recommendations as opposed to when personal data is used to drive insights or when data-driven modeling is used to drive real-time decisions for maintaining the health of a grid are vastly different in terms of consequences. This dissertation looks at analytical uncertainty in two real-world scenarios: i) how sensitive information leakage can be prevented during the open data release process with data custodians being the stakeholders, and ii) how errors in energy forecasting can be detected or prevented when deploying them in power systems, with grid operators being the stakeholders. Across both these scenarios, this dissertation investigates how interactive visualization workflows can empower respective data stakeholders to reveal privacy vulnerabilities in open datasets and improve trust in AI forecasting models within the power sector. The first contribution is a systematic analysis of existing visual analytics methods for addressing data privacy and examining research gaps and future opportunities. Building on this foundation, an ethical hacking exercise was conducted to identify vulnerabilities in the open data ecosystem, leading to the second contribution of this dissertation: the development of the PRIVEE workflow, which enables data defenders to assess disclosure risks associated with open datasets. This dissertation showcases the effectiveness of PRIVEE through case studies in collaboration with domain experts. Recognizing the need to understand the utility of linked datasets,

the third contribution presents the algorithm for a utility metric and the VALUE interface, allowing users to explore the utility of joining datasets across over 100 open data portals. This can quickly escalate into a combinatorial explosion due to the various factors involved in joining multiple datasets differently. Thus, as the fourth contribution, this dissertation explores how visual analytic interventions can help balance privacy and utility factors in the context of multi-way joins through the web-based interface LinkLens. Finally, the dissertation extends these principles to the energy sector, contributing to the development of the Forte application, which helps grid operators evaluate AI model performance. This work enhances human-data trust and informed decision-making by equipping stakeholders across disparate domains with interactive visualization workflows.

INTERACTIVE VISUALIZATION WORKFLOWS FOR MITIGATING ANALYTICAL UNCERTAINTY

by
Kaustav Bhattacharjee

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Data Science Computing Option

Department of Data Science

December 2024

APPROVAL PAGE

INTERACTIVE VISUALIZATION WORKFLOWS FOR MITIGATING ANALYTICAL UNCERTAINTY

Kaustav Bhattacharjee

Dr. Aritra Dasgupta, Dissertation Advisor — Date
Assistant Professor, Department of Data Science, NJIT

Dr. Chase Wu, Committee Member — Date
Professor, Department of Data Science, NJIT

Dr. Mengnan Du, Committee Member — Date
Assistant Professor, Department of Data Science, NJIT

Dr. Salam Daher, Committee Member — Date
Assistant Professor, Department of Informatics, NJIT

Dr. Soumya Kundu, Committee Member — Date
Staff Research Engineer, Pacific Northwest National Laboratory, Richland, WA

BIOGRAPHICAL SKETCH

Author: Kaustav Bhattacharjee

Degree: Doctor of Philosophy

Date: December 2024

Undergraduate and Graduate Education:

- Doctor of Philosophy in Data Science,
  New Jersey Institute of Technology, Newark, NJ, 2024
- Bachelor of Technology in Information Technology,
  West Bengal University of Technology, Kolkata, India, 2016

Major: Data Science Computing Option

Publications:

K. Bhattacharjee, S. Kundu, I. Chakraborty, and A. Dasgupta, "Who should I trust? A Visual Analytics Approach for Comparing Net Load Forecasting Models," *IEEE Power & Energy Society Grid Edge Technologies Conference & Exposition (Grid Edge): San Diego, USA: IEEE, 2025* (In Publication)

K. Bhattacharjee, S. Kundu, I. Chakraborty, and A. Dasgupta, "Forte: An Interactive Visual Analytic Tool for Trust-Augmented Net Load Forecasting," *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT): Washington D.C., USA: IEEE, 2024*

J. Yuan, K. Bhattacharjee, A.Z. Islam, and A. Dasgupta, "TRIVEA: Transparent Ranking Interpretation using Visual Explanation of Black-box Algorithmic Rankers," *The Visual Computer,* vol. 40, no. 5, pp. 3615-3631, 2024

K. Bhattacharjee, and A. Dasgupta, "VALUE: Visual Analytics driven Linked data Utility Evaluation," *Workshop on Human-In-the-Loop Data Analytics (HILDA): Seattle, USA: ACM, 2023*

K. Bhattacharjee, and A. Dasgupta, "Power to the Data Defenders: Human-Centered Disclosure Risk Calibration of Open Data," *Symposium on Usable Security and Privacy (USEC): San Diego, USA: NDSS, 2023*

K. Bhattacharjee, A.Z. Islam, J. Vaidya, and A. Dasgupta, "PRIVEE: A Visual Analytic Workflow for Proactive Privacy Risk Inspection of Open Data," *Symposium on Visualization for Cyber Security (VizSec): Oklahoma City, USA: IEEE, 2022*

K. Bhattacharjee, M. Chen, and A. Dasgupta, "Privacy-Preserving Data Visualization: Reflections on the State of the Art and Research Opportunities," *Computer Graphics Forum,* vol. 39, no. 3, pp. 675-692, 2020

Presentations:

K. Bhattacharjee and A. Dasgupta, "Look before you Link: Interactive Visualization Workflows for Assessing Privacy-Utility Trade-offs in Linkable Open Data," *Poster Presentation, NJ Big Data Alliance Annual Symposium*, New Brunswick, New Jersey, USA, 2024

K. Bhattacharjee, "Do you trust the model? Interactive Visual Analytics for Trust Augmented Net Load Forecasting," *Oral Presentation, Research Symposium, Pacific Northwest National Laboratory*, Richland, Washington, USA, 2024

K. Bhattacharjee and A. Dasgupta, "Look before you Link: Interactive Visualization Workflows for Assessing Privacy-Utility Trade-offs in Linkable Open Data," *Poster Presentation, NYC Privacy Day, NYU*, New York City, New York, USA, 2024 (Best Poster Award)

K. Bhattacharjee, "Look before you Link: Privacy Risk Inspection of Open Data through a Visual Analytic Workflow," *Lightning Talk, Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, Anaheim, California, USA, 2023

K. Bhattacharjee, "Look before you Link: Interactive Visualization Workflows for Assessing Privacy-Utility Trade-offs in Linkable Open Data," *Oral Presentation, Graduate Research Day, NJIT*, Newark, New Jersey, USA, 2023 (Best Presentation)

K. Bhattacharjee, A.Z. Islam, and A.Dasgupta, "Is Your Privacy Lost in Transition? Analyzing Transitive Disclosure Risks in Open Datasets," *Research Proposal, 6th Workshop on Technology and Consumer Protection (ConPro)*, San Francisco, California, USA, 2022

This dissertation is dedicated to
Maa, Baba, Didi, Wifey and family...

ACKNOWLEDGMENTS

I owe immense gratitude to the many people who have supported me throughout this dissertation journey. First and foremost, I am deeply thankful to my advisor, Dr. Aritra Dasgupta, for his exceptional guidance and constant encouragement. His expertise and belief in my abilities have been invaluable, and this work would not have been possible without his mentorship.

I would like to express my sincere gratitude to Dr. Chase Wu, Dr. Mengnan Du, Dr. Salam Daher, and Dr. Soumya Kundu for agreeing to serve on my dissertation committee. Their insights and feedback have been instrumental in shaping this work.

I am deeply grateful to the Department of Data Science for the opportunities provided during my doctoral journey, and especially to Janine for her consistent support. As a newly established department, it has already demonstrated a remarkable commitment to nurturing its students and promoting their growth. I am also grateful to the Department of Informatics for its support during the initial stages of my doctoral studies. I am especially thankful to Prof. Keith Williams for the opportunity to serve as a Teaching Assistant in his courses. Additionally, I want to extend my heartfelt thanks to Jacinta and Arleth for their unwavering support throughout this journey.

At this juncture, I would like to acknowledge the staff at Pacific Northwest National Laboratory for their steadfast support during my internship. I want to thank Dr. Kundu for his invaluable mentorship; his guidance provided me with clear direction throughout my internship. I would also like to thank Denise for always answering my questions with a smile.

I want to thank my colleagues for their camaraderie and intellectual engagement. A special shoutout to Vrushali and Jun for their encouragement and support with all things academic.

I would also like to thank Ishani Di for her constant support, which has helped me stay focused and grounded throughout this journey. I could not have completed this dissertation without the support of my friends, Smarth, Naren, Anisha, and others, whose unwavering encouragement, engaging conversations, and welcome diversions from research have been invaluable throughout this journey.

My heartfelt gratitude goes to my family for their unconditional love and support. I am forever indebted to my parents, Baba and Maa, for their wise counsel and sympathetic ear. My sister, Poulami, has always been a source of inspiration in my life. I began this journey fueled by your encouragement. A special shoutout goes to my wife, Komal; this work would not have been possible without her. I would also like to thank Kaushik Da for his encouraging words and support. I would like to once again express my gratitude to everyone in my life who has supported me throughout my research journey.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Introduction

Uncertainty, often defined as the "goodness" of a result, is an inherent aspect of data analysis, arising from multiple sources [2]. These include data quality issues, modeling assumptions, measurement errors, and sampling biases. For instance, inaccuracies, missing values, and biases in the data can introduce uncertainty. Similarly, the choice of analytical techniques and assumptions made during modeling can influence results, while measurement errors and sampling biases add further complexity. However, analytical uncertainty, in particular, stems from the analytical process itself. Decisions made during data exploration, such as feature selection and preprocessing, impact outcomes, as does the selection of models, which often carry inherent variability. Additionally, parameter tuning and the subjective interpretation of results can further amplify this form of uncertainty.

Analytical uncertainty becomes particularly risky when applied to decision-making. In today's data-driven world, its impact varies dramatically across different domains. While an incorrect movie recommendation may lead to minor disappointment, inaccurate analyses of personal data or energy forecasting can have far-reaching consequences for individuals and society at large. Moreover, if not communicated clearly, analytical uncertainty can lead to financial losses and erode public trust. A notable example occurred in 1997 when the U.S. National Weather Service (NWS) predicted that the Red River in Grand Forks would rise to 49 feet. Based on this forecast, local authorities constructed levees to withstand a flood of up to 51 feet. Unfortunately, the river ultimately rose to 54 feet, causing billions of dollars in damage and the loss of lives [3]. Had the NWS properly communicated

the uncertainty in their forecast (+/- 9 feet), much of this devastation could have been avoided. Thus, effective policy-making, grounded in a clear understanding of analytical uncertainty, is essential for safeguarding public trust and financial resources.

In this context, there is a famous saying that *data is the new oil and Artificial Intelligence (AI) is the new electricity* [4, 5, 6]. This reflects the immense potential of data and AI to shape decisions and policies that impact people's lives. Access to large-scale data for these insights, nonetheless, remains a complex challenge. A much-needed boost came with the U.S. Government's Open Government directive in 2009, followed by the signing of the Open Data Charter by G8 leaders in 2013, which accelerated the adoption of *open datasets* that are freely available for use, reuse, and redistribution [7, 8, 9]. Though these are generally anonymized before release, joining two anonymized datasets based on quasi-identifiers can lead to the disclosure of sensitive information. Moreover, accessibility to these datasets can be considered a double-edged sword. On the one hand, open data movement has enabled free access to these datasets through open data portals like NYC Open Data [10], Kansas City Open Data [11], and City of Dallas Open Data [12], democratizing access to hitherto proprietary data. On the other hand, inadvertent data leaks could compromise the privacy of human data subjects. For example, in 2016, the Australian Department of Health released *de-identified* medical records for 2.9 million patients (10% of the population). Yet, researchers were able to re-identify the patients and their doctors using other open demographic information within a few months [13]. In another example, passengers' private information was disclosed through the public transportation open data released by the city municipal of Riga, Latvia [14]. These examples demonstrate that privacy, a fundamental human right, can be compromised when uncertainty in the analytical process of disseminating open datasets is not adequately addressed.

On the other hand, AI has a significant impact on our daily lives in many ways. Recent advances in artificial intelligence have accelerated the development of models across various domains, including healthcare, economics, politics, and smart grids. For example, Google's DeepMind Health has developed an AI system that analyzes retinal images to detect early signs of diabetic retinopathy, while the OECD uses artificial intelligence models to forecast weekly GDP growth using data from 46 countries [15, 16]. In another example, Pacific Northwest National Laboratory (PNNL) has developed DeepGrid, an open-source platform that uses deep reinforcement learning to help power system operators in creating more robust emergency control protocols for the electric grid [17]. Nonetheless, the uncertainty inherent in the results of these AI models can significantly impact people's lives. For example, inaccurate forecasting of power consumption in a region can lead to substantial losses for electric utility companies, costs that ultimately fall on consumers. US Department of Energy (DOE) also warns that AI models used to make predictions about unexpected events, such as extreme weather, "can lead to unpredictable or inaccurate behavior," potentially resulting in inefficient power restoration efforts or misallocation of resources by utility companies [18]. In this context, mitigating analytical uncertainty while interpreting results from AI models is essential in today's world.

In this context, visual analytics can play an important role in mitigating uncertainty in the analytical process. As Sacha et al. noted, it can provide interactive tools that allow users to explore and understand the propagation of uncertainties through the analytical pipeline [19]. By visualizing uncertainties at different stages of data transformation and analysis, visual analytics enables users to build awareness of potential sources of error or bias. This increased awareness can lead to more informed decision-making and help calibrate users' trust in the analytical outcomes. Furthermore, visual analytics can support trust building by offering transparency

in the analytical process, allowing users to interactively explore different scenarios and understand how uncertainties impact final results. By bridging the gap between machine-generated uncertainties and human trust-building processes, visual analytics can provide a framework for more reliable knowledge generation in complex analytical tasks. It can play a crucial role in both privacy-preserving data analysis and net load forecasting by helping stakeholders understand, analyze, and mitigate analytical uncertainty. In the domain of privacy-preserving data analysis, visualization tools can assist data owners and custodians in evaluating the effectiveness of different anonymization techniques (such as k-anonymity, l-diversity, and t-closeness) and understanding the trade-offs between data utility and privacy protection. For example, Rode et al. demonstrated how visualization can help in assessing disclosure risks and configuring appropriate levels of anonymization [20]. Montemayor et al. further showed how visual representations can aid in understanding privacy risks in complex datasets [21]. In the context of net load forecasting, interactive visualization can empower energy scientists and grid operators to explore net load variability, assess forecast errors, and analyze the effects of various input variables on model performance across different time periods and seasons. Dasgupta et al. showed that visual analytics can significantly enhance trust in model outputs during complex sense-making tasks, which is particularly relevant for understanding and interpreting net load forecasts [22]. Kandakatla et al. argued that these techniques would play a critical role in enabling trust-augmented artificial intelligence and machine learning (AI/ML) applications in the energy sector [23]. Furthermore, interactive visualizations can help in understanding the impact of increasing solar energy penetration on traditional forecasting models and in evaluating the reliability and robustness of deep learning models in real-world scenarios with noisy inputs. By providing interactive and intuitive interfaces, these tools can facilitate the exploration of complex datasets and model outputs, enabling stakeholders to make more informed

decisions in both privacy preservation and energy planning. Ultimately, these approaches can lead to more effective privacy policies, improved anonymization techniques, and more accurate and reliable net load forecasts, addressing the analytical uncertainties inherent in both domains.

Thus, in order to develop visual analytic solutions for analytical uncertainty mitigation, we first conducted a survey of privacy-preserving data visualization [24]. We categorized the existing literature into different visual analytic tasks and reflected on the research gaps and future opportunities. Some of the gaps thus identified revolved around uncertainty visualization, dynamic visualization of risks for privacy stakeholders, and privacy-aware citizen science. During this research, we recognized that open datasets are a key component in enabling open governance, thereby fostering trust between citizens and their government. To explore this further, we collaborated with domain experts to conduct an ethical hacking exercise focused on the open data ecosystem, aiming to identify vulnerabilities and assess the associated risks for different data stakeholders. Although these datasets are anonymized before release, we found multiple examples where joining two datasets could disclose sensitive information about the data subjects [25]. But finding these datasets is akin to finding a needle in a haystack, due to the combinatorial explosion that leads to uncertainty in the analytical process. This ethical hacking exercise allowed us to adopt an attacker's perspective and led to the development of PRIVEE, a risk inspection workflow that enables data defenders to inspect all possible combinations of their datasets and evaluate possible disclosures at the record level [26]. We also developed a web-based interface grounded in this workflow and in this dissertation, we discuss the visual analytic interventions required to perform the workflow through our interface. Additionally, we present two case studies that demonstrate the effectiveness of this approach in evaluating disclosure risks and inspecting actual disclosures.

During this research, we realized that users are also interested in understanding the utility of joining open datasets and whether doing so would be worthwhile. To address this, we developed a utility metric that helps users evaluate multiple join combinations and select the one most useful for their needs [27]. We also developed a web-based interface VALUE, where users can join datasets from over 100 open data portals and compare them based on the utility of the joined datasets. In this dissertation, we discuss the visualization techniques used to implement this interactive interface, along with a usage scenario. Through discussions with the research community, we learned that researchers often need to join multiple open datasets for their work. This, in turn, introduces greater combinatorial complexity due to the various factors involved in these multi-way joins. To address this, we developed the workflow LinkLens, which balances privacy and utility considerations in multi-way joins, helping researchers make informed decisions during the analytical process. In a similar fashion, we distill this workflow through a visualization interface and discuss its design principles through this dissertation.

As mentioned earlier, analytical uncertainty arises from the analysis process itself and can significantly impact the interpretation of results from AI models. To address this, we wanted to determine whether visual analytic solutions could enhance the interpretability of these models. Hence, in this dissertation, we collaborated with energy scientists to develop Forte, a visual analytics-based application that addresses analytical uncertainty in the results of net load forecasting models across diverse time periods and input scenarios [28]. Our system enables researchers and grid operators to assess net load variability, analyze the effects of various input variables on model performance, and evaluate forecast uncertainties in the presence of noisy inputs. Forte provides a broad understanding of various aspects related to net load forecasting, allowing users to compare model forecasts with actual net load values across different seasons and prediction horizons, and gain insights

into the impact of variables like temperature, humidity, and apparent power on net load forecasts. However, trust in the analytical process is crucial for the effective deployment and utilization of these models. Therefore, we extended this workflow to facilitate comparison among multiple models and enhance trust in the model outcomes [29]. This interface, incorporating carefully selected visual analytic interventions, facilitates the comparison of multiple models across various parameters, including solar penetration levels, dataset resolutions, and different hours of the day. By enabling users to compare our net load forecasting model with a reference model, we provide a comprehensive framework for evaluating model performance and building confidence in the results. Ultimately, these approaches enhance trust and confidence in net load forecasting models, supporting data-driven decision-making in energy planning and grid operations.

This dissertation addresses the critical challenge of analytical uncertainty in high-stakes decision-making scenarios, focusing on open data release and energy grid management. By developing interactive visualization workflows and tools tailored to the specific needs of data custodians and grid operators, we aim to empower stakeholders to make more informed decisions in situations where errors can have significant consequences. These tools are designed to be both effective and practical for real-world implementation, enabling users to better understand and mitigate uncertainty in their respective domains. This work takes a process-centric and stakeholder-first perspective, recognizing that the cost of incorrect decisions varies drastically depending on the context - from relatively benign movie recommendations to potentially catastrophic outcomes in power grid management. In the realm of open data, we investigate how sensitive information leakage can be prevented during the release process, with data custodians as key stakeholders. For energy forecasting, we explore methods to improve trust when deploying AI models in power systems, with grid operators as the primary stakeholders. These two scenarios were selected

based on the analytical pipeline described in Sacha et al.'s paper, where uncertainty can arise from any part of the analytical pipeline, be it the model/system side or the human user side [19]. Using this pipeline, we selected two decision scenarios: one from the human side (uncertainty while analyzing open datasets for disclosures) and one from the system side (uncertainty in a net load forecasting model). Ultimately, this work contributes to improving decision-making processes in high-stakes environments, enhancing privacy protection in open data, and building trust in AI forecasting for critical infrastructure like power grids.

In this dissertation, we begin by outlining the background along with the core concepts related to this problem. We then present a survey and analysis of the research gaps in this domain (Chapter 2). Following this, we highlight some of the vulnerabilities we observed in the open data ecosystem that could lead to the disclosure of sensitive information (Chapter 3). Building on this foundation, we introduce our disclosure risk inspection workflow, PRIVEE, and discuss the design principles used in this interface (Chapter 4). Next, we present the algorithm for our utility metric and discuss a usage scenario in which the users can explore the utility of joining open datasets using our interface, VALUE (Chapter 5). We extend these algorithms and design principles for multi-way joins and discuss how the privacy and utility factors can be balanced using our tool, LinkLens (Chapter 6). Following this, we discuss our tool, Forte, which helps grid operators and power scientists evaluate the performance of an AI model amid uncertainties arising from various input scenarios and noisy conditions (Chapter 7). Next, we discuss extending this workflow to allow model comparison across different time points and seasons, ultimately enhancing trust in the AI model (Chapter 8). Finally, we conclude by exploring how interactive visualization workflows can mitigate analytical uncertainty in both domain-agnostic and domain-specific use cases while also discussing potential future research directions (Chapter 9).

1.2 Background

In this section, we introduce some key concepts related to this dissertation. We begin by explaining uncertainty and its various types, followed by a brief overview of open datasets. Next, we examine the stakeholders in the open data ecosystem and their roles in managing uncertainty and safeguarding privacy. We then explore the re-identification problem in open datasets and review some of the anonymization methods proposed to address it. Following this, we shift focus to net load forecasting, discussing how uncertainty in analyzing AI model results can lead to misinterpretation. We conclude this section by exploring the role of human factors in visualizations, thus providing a foundation for privacy-preserving data visualizations that help mitigate analytical uncertainty.

What is Uncertainty?: Uncertainty in measurement is defined as the parameter associated with the result of a measurement that characterizes the dispersion of values that could be reasonably attributed to the quantity being measured (also known as measurand) [30]. For example, standard deviation can be used to measure the deviation of a variable along its mean. Similarly, a confidence interval is used to quantify the uncertainty surrounding a forecast or prediction. It shows how an estimate differs from the true value, helping users gauge the degree of confidence in the results [31]. In simpler terms, uncertainty is the measure of the "goodness" of the results of a process [2]. Several metrics are available to quantify uncertainty. For example, standard error and the coefficient of variation are commonly used to interpret economic estimates across different sectors. Standard error helps assess possible sampling error, while the coefficient of variation provides the relative standard error in comparison to the actual estimate. For instance, the UK's Office of National Statistics reported that the total turnover in the education sector for 2016 was £42,649 million, with a standard error of £526.8 million, yielding a 1%

coefficient of variation [32, 33]. Other metrics, such as confidence intervals and statistical significance, are also used to measure uncertainty.

Analytical uncertainty refers to the uncertainty in the results of an analysis, which can influence the decisions made based on those results. In short, it is the uncertainty in the process itself. This uncertainty can stem from multiple factors, including the data, the assumptions made, and the nature of the analytical questions posed. It is central to decision-making and risk analysis, as decision-makers need to grasp the uncertainty surrounding the impacts of their choices. Reducing analytical uncertainty requires decision-makers to align on the questions being asked and how the outcomes will inform their decisions. For instance, consider a scenario where a pharmaceutical company is evaluating the safety of a new drug. While the company may be interested in determining if the drug is effective, it is equally important to assess the degree of uncertainty surrounding potential adverse side effects. A misjudgment in this case could lead to serious health consequences for patients. Analytical uncertainty can be classified in several ways, with a common framework dividing it into "known knowns," "known unknowns," and "unknown unknowns" [34]. Known knowns (or aleatory uncertainty) refer to things we are aware of, such as the range of variability inherent in a model's prediction due to its probabilistic nature. This type of uncertainty is quantifiable but often unavoidable, though it can be mitigated through techniques like data smoothing. Known unknowns (or epistemic uncertainty) are things that we know we don't know and encompass the gaps in knowledge about system complexities, such as uncertainty about whether a model will perform consistently with noisy data. This can usually be quantified through sensitivity analysis and can be reduced by gathering more knowledge and filling those gaps. Unknown unknowns are things that we don't know we don't know. It arises from factors that were previously unrecognized and thus cannot typically be quantified. One example is the analytical uncertainty in joining open

datasets and the risk of revealing sensitive information. Our work aims to develop visualization workflows to enable users to systematically explore these open dataset joins, identify vulnerabilities, and take corrective action. The aim of this dissertation is to convert "unknown unknowns" into "known unknowns" by equipping users with tools to evaluate uncertainty within the open data ecosystem. In addition, this research investigates whether visualization workflows can also assist with known unknowns. To that end, we developed visual analytics tools to address uncertainty in net load forecasting, a domain currently classified under known unknown uncertainty. Researchers have explored various ways to assess the performance of AI models, and this dissertation aims to enhance these methods with our visual solutions. However, addressing the disclosure of sensitive information through open datasets remains a significant challenge, as it still resides in the unknown unknown category. This underscores the need to first understand the open data ecosystem and its stakeholders before defining processes and workflows to tackle this issue.

Open data ecosystem: The journey of the open data ecosystem commenced with the US Government's Open Government Directive in 2009, which mandated federal agencies to make their data publicly available [8]. This initiative sought to enhance transparency, participation, and collaboration by increasing access to federal datasets. In 2013, this movement gained international traction when G8 leaders signed the Open Data Charter, committing to principles such as making government data open by default and improving its quality, accessibility, and re-usability [9, 35]. As a result, open datasets, covering areas like education, health, transportation, and crime statistics, are now made available without restrictions by government agencies, research institutions, and public organizations. Platforms like Data.gov [36], the London Datastore [37], and NYC Open Data Portal [38] facilitate easy access to these datasets for public use. The breadth of data spans public administration, environmental monitoring, scientific research, and economic indicators. Nowadays,

tools like Urban Profiler [39], Socrata Discovery API [40], Urban Forest[41] have made it even easier to find relevant data for research. It has also proven instrumental in shaping public policies, as exemplified by the Behavioral Risk Factor Surveillance System (BRFSS) in the United States, which has been used to monitor and respond to public health emergencies in real-time, such as developing the public health response to Hurricane Katrina in 2005 and tracking H1N1 vaccine uptake during the 2009 influenza pandemic [42]. In urban planning, cities like New York have leveraged open transportation data to optimize public transit routes and reduce congestion [43, 44]. Through such applications, open data has not only enhanced policy effectiveness but also fostered civic engagement and economic development.

Who are the data stakeholders?: In Figure 1.1, we illustrate the different types of stakeholders and their roles in the context of privacy-preserving data visualization in open datasets. The stakeholders with the highest responsibility in this ecosystem are the data owners, who collect and have proprietary rights over the collected data, and the data custodians, who have the responsibility of enforcing policies and safeguarding the privacy of the data. Cambridge Analytica's much-debated and questionable use of Facebook data [45] demonstrates how privacy preservation responsibilities can be misused. Data subjects are the individuals (e.g., people on Facebook) who provide implicit or explicit consent to different agencies for collecting their personal data. They need to be cognizant of the risks of sharing personal data and understand the privacy policies of companies, a task that is often complex and inconvenient. In fact, recent studies have demonstrated the lack of effectiveness of privacy policies of online companies [46], and even worse, the deliberate use of dark patterns for subverting policy implementations [47]. Data consumers are analysts or the general public with appropriate levels of access to sanitized data who want to derive insights without violating privacy. In many cases, data subjects themselves are consumers (e.g., patients mining electronic health records and people trying to understand trends

Stakeholder	Role in the data ecosystem	Stake for privacy
 **Data owner**	An entity which owns data about people or individuals whose data is captured. Examples: hospitals, social media companies, social media users	• Wants to understand risks of releasing data for public consumption • Implement privacy legislation in the form of policies
 **Data subject**	Individuals whose data are represented in databases or are collected by applications. Examples: patients, common public	• Decide whether or not trust an agency for collecting their data • Understand implications of privacy policies
 **Data custodian**	An entity with credentials for accessing a private database or a 3rd party entrusted with data analysis. Examples: Cambridge Analytica	• Have access to the original or a limited version of the data • Implement privacy legislation in the form of policies
 **Data consumer**	Any person who is the intended audience for shared data or analysis. Examples: data analysts, scientists, policy makers, and the public	• Access anonymized data • Derive value from data without getting to know sensitive information
 **Attacker**	Anyone with the goal of breaching privacy and knowing about people. Examples: any attacker with or without background knowledge about the collection	• Link publicly and privately available information with the intent of privacy breach • May or may not have background knowledge about individuals in a database

Figure 1.1 Different stakeholders in the open data ecosystem from a privacy perspective: Data owners and custodians need to preserve and protect the privacy of data subjects (i.e., individuals represented in a dataset) from insider or outside attackers. Privacy-preserving visualization is used by data owners or custodians for understanding privacy-utility trade-offs and is also used by data subjects, who want to understand privacy policies, and data consumers, who want to derive value from anonymized data.

in survey data). Attackers are people or enterprises with malicious intent constantly attempting to breach private databases or attack privacy-preservation mechanisms duly enforced in publicly available data. They can try to re-identify anonymized records by linking multiple open datasets.

Re-identification via linking: When releasing data, merely suppressing personally identifiable information (PII), like name, social security number, email address, etc., is necessary yet not sufficient. *Quasi-identifiers* [48], like age, gender, zip code, etc., can be exploited by attackers to breach privacy by linking attributes from publicly available data sources (e.g., voter registration data) and privately accessible information (e.g., hospital data or web access data). This is popularly known as the *data linking* problem [49], and various data anonymization methods [50] like generalization, suppression, perturbation, clustering, etc., are used to tackle this problem. These methods typically produce an anonymized static data table, a modified data mining algorithm, or an anonymized visualization. Most of these methods constitute the non-interactive setting of privacy-preservation, where, once released, the data owner does not have any control over the data or the mining results, and the drawbacks of such a "release-and-forget" model [51] have been questioned by recent studies. Next, we will discuss the anonymization methods followed by the implications of this release-and-forget model.

Anonymization methods: One of the most widely used anonymization methods is the k-anonymity model. It states that a dataset is k-anonymous if the information for each record in the dataset cannot be distinguished from at least $k-1$ other records [52, 53]. For example, if $k = 3$, then a k-anonymized dataset will have at least 3 similar combinations for each record of potentially identifying variables. But k-anonymity does not provide a guarantee against attackers having background knowledge or homogeneous attacks.

Company	Position	Nationality	Zip	Age	Disease
Alpha	Director	Japanese	10001	32	Galactosemia
Beta	Manager	Indian	11049	53	Cancer
Gamma	Associate	American	10011	38	Galactosemia
Beta	Manager	Russian	10004	43	Fatty Liver
Alpha	Manager	Japanese	10014	48	Hepatitis B
Delta	Consultan	Indian	10017	34	Galactosemia
Gamma	Associate	American	11042	57	Hepatitis B
Delta	Manager	American	10007	42	Hepatitis B
Gamma	Director	Japanese	11043	51	Galactosemia
Beta	Manager	Russian	10009	35	Galactosemia
Delta	Associate	Indian	10019	42	Fatty Liver
Gamma	Manager	Japanese	11047	63	Fatty Liver

Table 1: Original dataset

Company	Position	Nationality	Zip	Age	Disease
*	*	*	100**	<40	Galactosemia
*	*	*	100**	<40	Galactosemia
*	*	*	100**	<40	Galactosemia
*	*	*	100**	<40	Galactosemia
*	*	*	110**	>=50	Galactosemia
*	*	*	110**	>=50	Cancer
*	*	*	110**	>=50	Hepatitis B
*	*	*	110**	>=50	Fatty Liver
*	*	*	100**	4*	Hepatitis B
*	*	*	100**	4*	Fatty Liver
*	*	*	100**	4*	Fatty Liver
*	*	*	100**	4*	Hepatitis B

Table 2: k-anonymous dataset (k=4)

Company	Position	Nationality	Zip	Age	Disease
*	*	*	1000*	<50	Galactosemia
*	*	*	1000*	<50	Fatty Liver
*	*	*	1000*	<50	Hepatitis B
*	*	*	1000*	<50	Galactosemia
*	*	*	1104*	>=50	Hepatitis B
*	*	*	1104*	>=50	Galactosemia
*	*	*	1104*	>=50	Fatty Liver
*	*	*	1104*	>=50	Cancer
*	*	*	1001*	<50	Galactosemia
*	*	*	1001*	<50	Hepatitis B
*	*	*	1001*	<50	Galactosemia
*	*	*	1001*	<50	Fatty Liver

Table 3: l-diverse dataset (l=3)

Figure 1.2 Examples of data anonymization based on the k-anonymity and l-diversity metrics: k-anonymity ensures sufficient group size (here k =4) so that an individual cannot be distinguished within that group and l-diversity ensures sufficient diversity in the values of an attribute (here, l=3), so that the exact values of a sensitive attribute cannot be detected from this dataset.

Let us refer to a dataset as shown in Figure 1.2. Figure 1.2 (Table 1) represents a dataset from clinical records, and Figure 1.2 (Table 2) is its 4-anonymized version. Suppose we know that John is an American associate of age 38 living in the zip code 10011, then we can easily decipher from Figure 1.2 (Table 2) that he has Galactosemia. This is the problem of homogeneous attacks. Again, suppose we know that Kabir is a 42-year-old Indian associate who lives in zip code 10019 and works for the company Delta. In that case, we can easily say he has either Hepatitis B or Fatty Liver. But if we have background knowledge (e.g., associates of the company Delta have been immunized against Hepatitis B), we can infer that Kabir has Fatty Liver. Thus, these types of attacks cannot be prevented even if the dataset is k-anonymized.

This problem is addressed by another anonymization method, the l-diversity model [54], which guarantees sufficient diversity in the value of attributes. The data from Figure 1.2 (Table 1) can be represented in a 3-diverse way in Figure 1.2 (Table 3). Here, each block of four records has a minimum of three varieties of the disease each. Now even if we know that John is an American associate of age 38 living in the zip code 10011, we can only decipher that he has either Galactosemia, Fatty Liver or Hepatitis B. Also, if we know that Kabir is a 42-year-old Indian associate who lives in zip code 10019 and works for the company Delta, and we have the background

knowledge that the associates of the company Delta have been immunized against Hepatitis B, we cannot tell with a guarantee that he has Galactosemia or Fatty Liver. Hence, both the problems of k-anonymity can be avoided through l-diversity.

On the other hand, l-diversity has its own limitations. l-diversity may be difficult and unnecessary to achieve. For example, let us assume our data in Figure 1.2 (Table 1) contains only one sensitive attribute, i.e., whether the person has a disease or not (Yes/No), and has around $100,000$ records. 98% of them have a disease (Yes), and only 2% of them do not have any disease (No). In order to have a 2-diverse table, there can be, at most, 2000 equivalence classes. Moreover, l-diversity is insufficient to prevent attribute disclosure. In our previous example, suppose an equivalence class has 49 negative records and one positive record. This implies that any individual in this class will have a 98% possibility of not having a disease instead of the overall 2% in the whole dataset. This is called a skewness attack. Moreover, l-diversity is also not immune to similarity attacks. For example, in Figure 1.2 (Table 3), if someone belongs to the last equivalence class and knows that Galactosemia, Hepatitis B, and Fatty Liver are diseases related to the liver, then we can easily decipher that any individual belonging to that equivalence class has liver disease.

The above scenarios can be alleviated using the t-closeness [55], which measures the distance between the distribution of a sensitive attribute in an equivalence class and the distribution of the attribute in the whole table and guarantees that the distance is at most t. An even more robust and popular anonymization concept is that of differential privacy [56, 57]. Differential privacy guarantees the following: a) anyone analyzing the results of a differentially private analysis will make the same inference about an individual's private information, irrespective of the fact whether the individual's private data was used in the analysis or not [58] and b) privacy protection against a gamut of privacy attacks, including linkage attacks, reconstruction attacks, and differencing attacks [56]. But protecting the data subjects'

privacy can impact a dataset's utility to some extent. The choice between the privacy and utility preservation of datasets can be made by a human expert, and visual aids can help the expert make this decision.

Net load forecasting: Net load, also known as residual load, refers to the difference between total electricity demand and the electricity generated by variable renewable energy sources, primarily wind and solar power [59]. Forecasting net load is vital for grid operators and utilities to maintain a stable and reliable power supply, as it helps them anticipate the amount of conventional generation needed to meet demand. This forecasting process involves predicting both the total electricity demand and the expected renewable energy generation, then calculating the difference between these two values. Accurate net load forecasting is essential for efficient grid management, as it enables operators to optimize the dispatch of conventional power plants, manage energy storage systems, and implement demand response programs effectively. The importance of net load forecasting has grown significantly in recent years due to the increasing penetration of renewable energy sources in power systems worldwide. This trend also presents challenges for forecasters, particularly with the rise of behind-the-meter solar installations, which can introduce uncertainties in both demand and generation predictions. High levels of distributed solar generation can lead to more volatile net load profiles and create phenomena like the "duck curve," where rapid ramps in net load occur during sunset hours, necessitating more sophisticated forecasting techniques and flexible grid operations [60]. Artificial Intelligence (AI) is playing a crucial role in improving net load forecasting by leveraging advanced techniques such as Long Short-Term Memory (LSTM) networks, which can capture complex temporal patterns and spatial interactions in the distribution grid [61].

Human Factors and the role of visualization: As apparent from the above discussion, several human factors are involved with all stages of privacy-preservation of data, be it the choice of anonymization methods, evaluation of trade-offs, or the

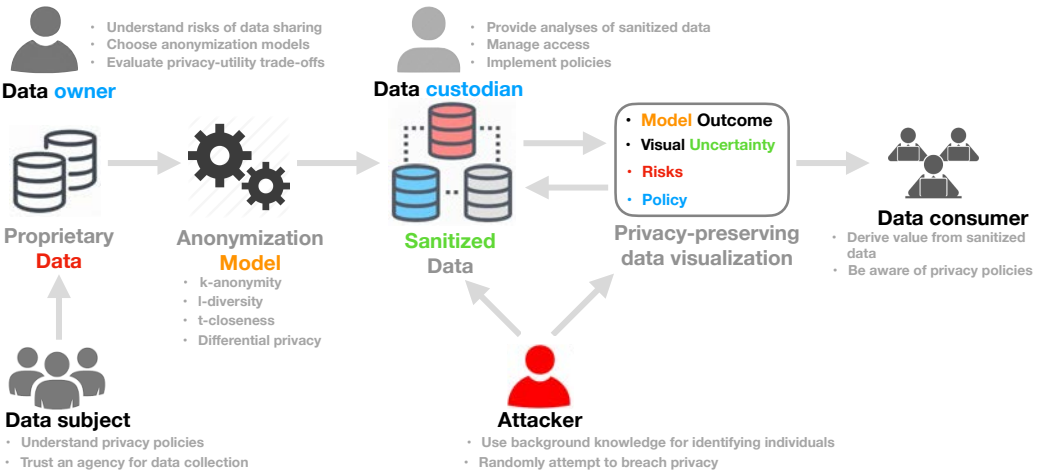


Figure 1.3 Data flow and roles of stakeholders: Privacy-preserving data visualization involves visual representation of outcomes of different anonymization models, addition of visual uncertainty as defense mechanism, evaluation of disclosure risks, and visualization of policy implications. The abiding goal in all of these cases is to guarantee a minimum level of privacy that can protect the data with respect to attack scenarios.

various attack scenarios, often triggered by attackers' background knowledge [62, 63]. This is illustrated in the privacy-preserving data visualization pipeline shown in Figure 1.3. Data owners often need to control access to proprietary data and protect it from even insiders in a company, and therefore visualization can help them understand the risks [20, 21] and more transparently configure appropriate levels of anonymization and data accessibility. Disclosure risk minimization [64, 65] is a key goal for both data owners and data custodians, particularly when outside adversaries can mine the released data or the results of the analysis process by using their background knowledge. Visualization can help understand the privacy guarantees and risk-utility trade-offs. For data consumers, a better understanding of mental models of personal privacy [66, 67] can let us know what kind of human inputs and interaction mechanisms should be considered for developing visualization interfaces. In the case of net load forecasting, the grid operators, analysts, and policymakers essentially act as data consumers, interpreting complex forecasting models and making critical decisions based on their outputs. In our survey (Chapter 2), we aim to understand

whether state of the art in privacy-preserving data visualization addresses these known unknowns and, if so, what are the emerging trends, patterns, and gaps thereof.

CHAPTER 2

LITERATURE REVIEW

Privacy preservation has become an antithesis to the idea of a digital data-driven era. Be it the smart devices that we use, the online services we access, or even the places we visit, data about our activities, identity, habits, and preferences are being collected at an unprecedented rate. Privacy, a fundamental human right, is often considered collateral damage in a bid to personalize and monetize commercial services offered to people. Several researchers have recently posited that the data landscape is confronted with a privacy *crisis* [68, 69, 70], and to fix it, an immediate collaborative effort among multiple stakeholders in the data ecosystem is needed.

Who are these stakeholders? In the related research areas of privacy-preserving data publishing [71] and mining [72], there has been extensive discussion on the role of different stakeholders. The stakeholders with the highest responsibility in this ecosystem are the data owners, who collect and have proprietary rights over the collected data, and the data custodians, who have the responsibility of enforcing policies and safeguarding the privacy of the data. Cambridge Analytica's much-debated and questionable use of Facebook data [45] demonstrate how privacy preservation responsibilities can be misused. Data subjects are the individuals (e.g., people on Facebook) who provide implicit or explicit consent to different agencies for collecting their personal data. They need to be cognizant of the risks of sharing personal data and understand the privacy policies of companies, a task that is often complex and inconvenient. In fact, recent studies have demonstrated the lack of effectiveness of privacy policies of online companies [46], and even worse, the deliberate use of dark patterns for subverting policy implementations [47]. Data consumers are analysts or the general public with appropriate levels of access to

sanitized data who want to derive insights without violating privacy. In many cases, data subjects themselves are consumers (e.g., patients mining electronic health records, people trying to understand trends in survey data). Attackers are people or enterprises with malicious intent, who are always attempting to breach private databases or attack privacy-preservation mechanisms duly enforced in openly available data. While regulations such as HIPAA[73], or more recently, GDPR [74] aim to protect data subjects against such malicious attacks by enforcing strict regulations for releasing data, recent studies have demonstrated how even heavily anonymized datasets run the risk of privacy breach, where demographic attributes in openly available data can be used to re-identify about 99% of Americans [51]. The latter case study is a telling commentary on how static privacy-preservation mechanisms (where anonymized data is released without any subsequent checks of risks) are inadequate in the face of evolving threats and attack scenarios.

Given this rather bleak picture of privacy in the real world, our attempt in this state-of-the-art report is to: a) investigate if and how visualization can empower data owners, subjects, custodians, and consumers to have a transparent understanding of privacy implications and b) provide guidance on how visualization can play a significant role towards addressing the socio-technical dimensions of data privacy. In the process, we analyze how a futuristic research agenda can adapt to the needs of the different stakeholders. As illustrated earlier, people's roles define what kind of stake or incentives they have for preserving or breaching data privacy. For example, a biologist who runs a research lab or a company that collects data about people's social media interests, would want to get guidance on the risks of sharing data with a broader group of people. A data custodian, like Cambridge Analytica, needs to have checks and balances in place to ensure people's identities are not revealed due to the use of demographic data. Data consumers, like a social scientist trying to understand the correlation between demographics and economic indicators of a region, need to

derive value out of anonymized data and overcome the potential loss of value due to suppression or omission of sensitive information. With the ubiquitous availability of smartphones, data subjects are often at the receiving end of privacy violation as personal data is being collected at an unprecedented rate, often with dubious policies and purposes. In rare cases like the currently unfolding *COVID-19* pandemic, such data collection becomes a societal need for contact tracing [75], which also brings privacy risks in its wake and solutions [76] need to be developed where public health and individual privacy are not considered to be trade-offs in policy implementations.

Visualization can play a critical role in all these scenarios, as evidenced by the state-of-the-art literature on privacy-preserving data visualization. This field of research has imbibed and extended concepts from the privacy-preserving data publishing [77] and mining [78] communities for developing visualization-specific solutions for anonymization, controlled access, and utility and risk analysis of released datasets. Our goal in this survey is to take a problem and task-driven approach towards organizing the existing research. This approach is motivated by the fact that privacy is as much a computational challenge as it is a challenge related to consideration of human factors across domains like healthcare [79] and social networks [80, 81].

To study these factors, we introspect about the privacy problem and the related goals of stakeholders and then map those back to the anonymization methods and visualization techniques. Our survey makes three contributions: i) Task-driven understanding of the privacy preservation goals with regards to different application scenarios and multiple stakeholders in the data ecosystem, like the data owners, data custodians, and data consumers, ii) Comparison of tasks and techniques for privacy-preserving data visualization and a critique of the design space, and the iii) Analysis of gaps and emerging research opportunities by establishing the context of

privacy-preservation related challenges in the realms of both privacy-related research gaps and emerging research areas in visualization and visual analytics.

After discovering the vulnerabilities in the open data ecosystem, we decided to develop a workflow that helps inspect more such vulnerabilities and disclosures and thus helps to preserve the privacy of the data subjects mentioned in the open datasets. As discussed above, visual analytic interventions can help emulate this workflow through a web-based interface. Thus, in order to understand the application of privacy preservation in data visualizations, we have conducted a survey of the literature of this domain and classified it according to the anonymization methods used by them, along with the visualization tasks we identified and the associated techniques to implement them. We will first discuss the survey methodology and the classification scheme in detail, followed by the task and techniques and our analysis of the gaps and future opportunities in this domain.

2.1 Survey Methodology And Classification Scheme

In this section, we first describe our survey methodology. Specifically, we discuss the definition of privacy that is relevant for visualization and describe our analysis workflow.

2.1.1 Definition and scope for literature search

The field of privacy-preserving data visualization lacks a thorough characterization of human-specific needs and goals. Depending on whether the target user is a data owner, a data subject, or a data consumer, the uses of visualization are likely to be vastly different (Figure 1.3). We look at the relevance of visualization in privacy from the dual lens of input and output privacy [82, 83, 72], where input privacy involves the transformation of a dataset into its privacy-preserving form through anonymization methods, and output privacy involves judgment about the analysis outcomes of the privacy-preserving dataset: whether the analysis or the visualization

is also privacy-preserving, i.e., how difficult it is for an attacker to infer sensitive knowledge by observing the patterns.

Since privacy-preserving data visualization is a relatively newer research area as compared to other areas of visualization research, we wanted to collect papers that reflect both the theoretical and practical aspects of visualization usage in the context of privacy. To that end, we followed a three-stage process for paper collection. In the first stage, we performed a broad search on IEEE and ACM digital libraries and Google Scholar with various combinations of keywords such as "privacy and visualization", "privacy-preserving visualization", "privacy and visual", "privacy and human factors", etc. This phase gave us a data-driven idea of the domains in which we were most likely to find privacy-preservation techniques and strategies involving data visualization. The healthcare domain was the most frequent one we encountered through our initial exploration, with the social science domain being a distant second.

In the second stage, we performed a deeper search into top-ranked domain-specific journals from healthcare, such as the *Journal of Biomedical Research*, and social science, such as the *Social Science Journal*. We collected more than a hundred papers from them by repeating the search terms "privacy and visualization". We also looked into the Google Scholar citations of these papers. Our inclusion criterion was to consider any paper that proposes a visualization method or technique as part of their privacy-preservation theory and applications. Most social science papers did not satisfy this criterion and had to be excluded from our collection. For papers published in visualization-specific venues, we collected research papers related to privacy-preserving data visualization by focusing our search on leading visualization publications from the past twenty years. These include proceedings of the Information Visualization Symposium/Conference and journals such as *IEEE Transactions of Visualization and Computer Graphics (TVCG)*, *Computer Graphics Forum*, *ACM CHI Conference*, and *IEEE PacificVis Symposium*.

In the third stage of our paper search process, we considered publication venues such as *ACM CHI* and *ACM SOUPS*, from where we collected several papers related to visualization and privacy that were specific to the security domain or were domain agnostic. We applied the same inclusion criterion for these papers.

Before applying our inclusion criterion, our corpus comprised about 400 papers. We carefully checked our corpus even after applying the inclusion criterion and filtered out any paper that only reflected on a *potential* use of visualization or a *potential* breach of privacy in a dataset, without discussing any specific method or technique. We finally ended up with 38 papers with contributions in the visualization domain and the specific application domain (e.g., healthcare, social science, and security and privacy). The latter collection helped us take a user-centered approach which was our goal from the onset.

2.1.2 Classification scheme

We derived a classification scheme (Figure 2.1) to characterize the different research contributions in the literature. We look at the problem of privacy preservation from an end user's perspective and focus on whether the techniques, methods, or applications are designed for a data owner, data consumer, or data subject. Due to the inherent similarity of the roles of data owners and data custodians from the perspective of privacy preservation and also in the context of the work we surveyed, we treat them as one group of users. Data owners, who hold proprietary rights for the collected data (e.g., social media companies or hospitals), aim to anonymize the data, implement access control, and implement accountability in order to increase the levels of privacy preservation. On the other hand, data consumers (e.g., analysts using social media data, scientists using health-care data for research, laypeople using data from fitness trackers) are generally provided with an anonymized version of the data or the visualization for deriving value out of it. In our collection, we found there is

an even split between the techniques that consider these groups as their target users. Data owners must be cognizant of the risks owing to **identity disclosure** (i.e., data consumers knowing exactly who the individuals are, from the data points representing them) and **attribute disclosure** (i.e., data consumers knowing the value of different quasi-identifiers or sensitive attributes) risk scenarios. They also have to understand what kind of **attack scenarios** a released data or a visualization may be subjected to based on the availability of other data sources or the background knowledge of the attacker. Visualization systems themselves can be subject to attack, and thereby the privacy guarantees might be compromised [84]. When a person with a data owner's role in a company needs to share data internally, they also have to implement appropriate **access control** mechanisms: people with only certain roles and privileges can access de-anonymized versions of the data. Data consumers must overcome the barriers of anonymization to derive value from the data. When a consumer is subject to data collection (e.g., whenever we use services on our smartphones), they also need to be cognizant of the disclosure risks associated with sharing their information. One of the most critical challenges in information privacy is the trade-off between privacy and the value or utility of the data. We observed that while there is a systematic approach toward defining what privacy means and how anonymization methods can help achieve different levels of privacy, in comparison, there is a lack of consensus about how the utility of anonymized data or a visualization derived from it can be qualified or quantified. The trade-off between privacy and utility affects both data owners and consumers. Based on the choice of anonymization methods like k-anonymity, l-diversity, and t-closeness (as discussed in Chapter 1), the degree of reduced utility of the data will vary.

The privacy problems faced by data owners [85, 86, 87, 88] can be described as follows based on our collection:

- How to choose anonymization methods that minimize disclosure risks and maximize the utility of the shared data?
- How to develop a privacy-preserving interface or visualization which will help users leverage interactive capabilities without leaking information about sensitive attributes?
- What are the vulnerabilities faced during the data flow between organizations that may result in policy non-compliance?
- How to share data between different entities (sensors, people, etc.) without privacy leakage?
- What are the degrees of re-identification risks, based on external information or users' background knowledge, once the data or the visualization is publicly accessible?
- Can attack scenarios be predicted, and accordingly, how can defense mechanisms be integrated within an anonymized visualization?

While some of the above problems also affect data consumers, we can describe the additional privacy problems faced by data subjects and consumers based on the literature [89, 90, 91, 81, 92, 93, 94] as follows:

- How to assess one's privacy on Online Social Networks (OSN)?
- What are the permissions requested by mobile applications, and how is the shared information used?
- Does a website sell or misuse private information by stating them explicitly in the privacy policies? Can data consumers be better aware of potential dark patterns [47]?
- How can data owners and consumers communicate better through more interpretable privacy policies?

We use this categorization and problem definition to describe the visualization-specific tasks, solutions, and challenges addressed in the literature, which we describe in detail in the following sub-sections. This scheme has been illustrated with the examples from our corpus in Figure 2.1.

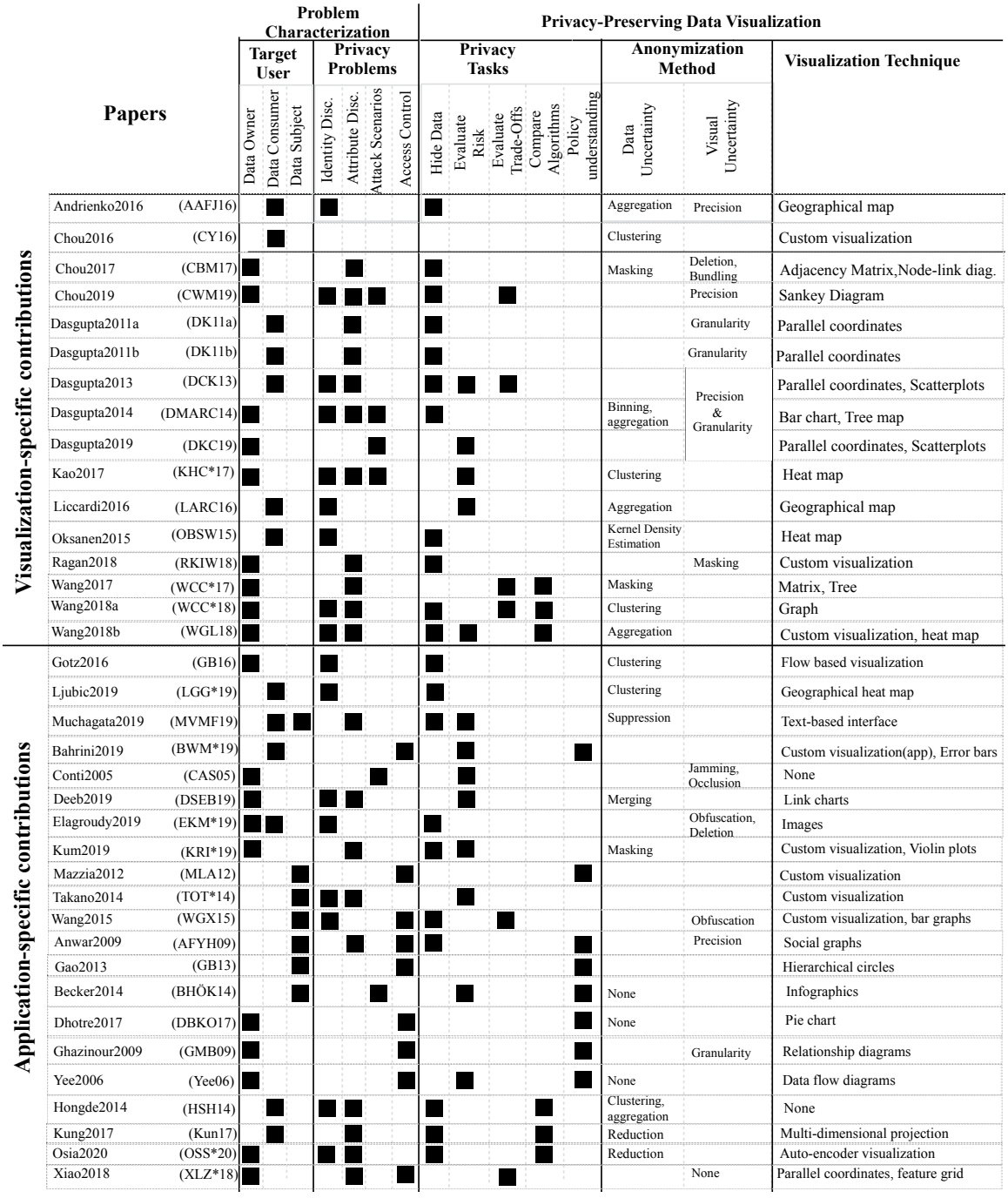
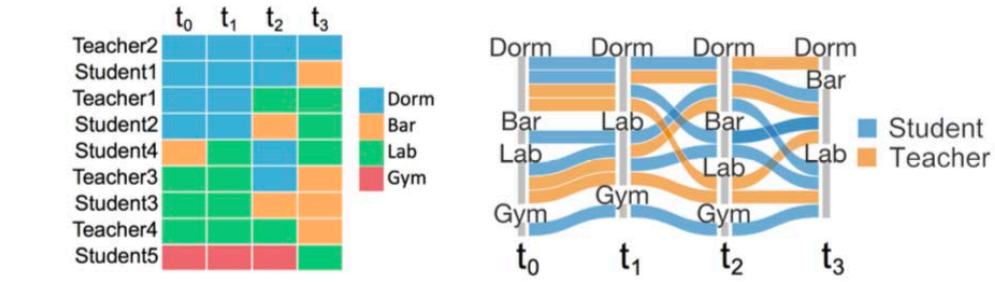
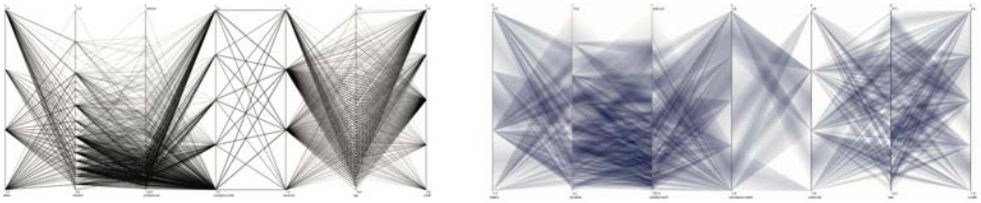


Figure 2.1 Classification Scheme for describing the literature on privacy-preserving data visualization: This scheme is based on the target users, privacy problems, visualization tasks intended to solve those problems, and the anonymization method used in conjunction with different visualization techniques.



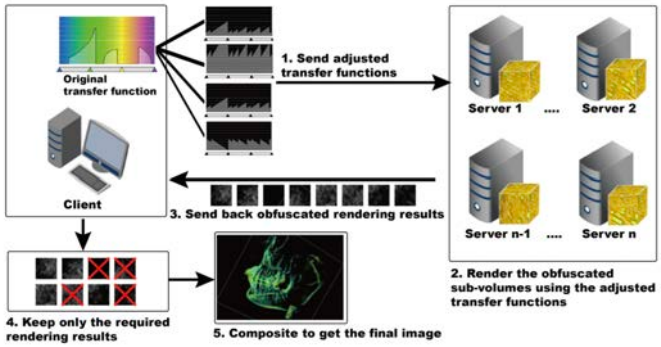
Data Uncertainty based anonymization: Sankey diagram used for transforming the visualization on the left to a privacy-preserving one

a [Chou2019]



Visual Uncertainty based anonymization: Clustered parallel coordinates on the right ensures a minimum level of privacy by using a combination of screen space metrics and anonymization methods

b [Dasgupta2011b]



Visual Uncertainty based anonymization: A volume rendering pipeline which uses obfuscation methods and customized transfer functions for generating a final image which guarantees a minimal level of privacy

c [Chou2016]

Figure 2.2 Illustrating anonymization methods: Based on data uncertainty and visual uncertainty.

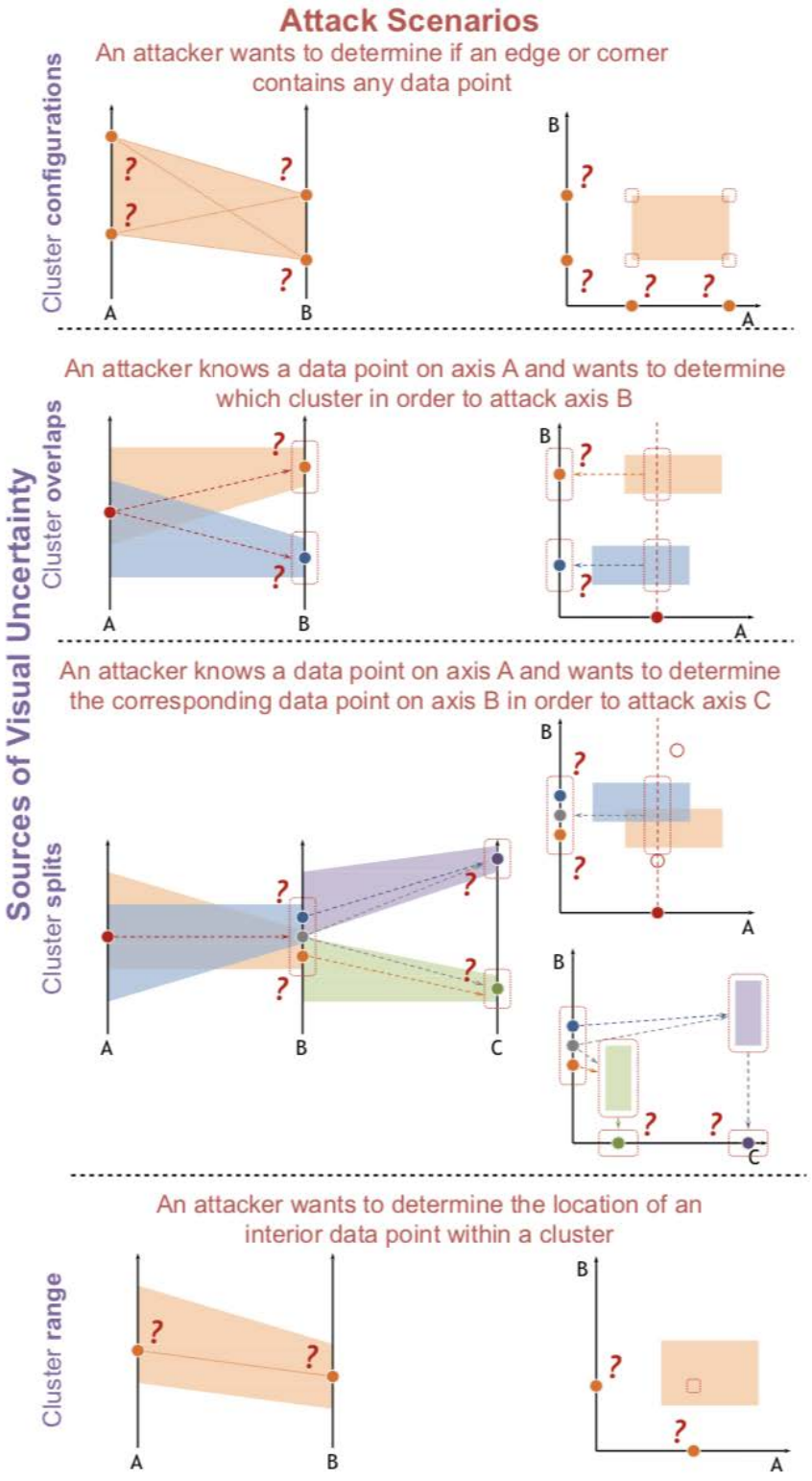


Figure 2.3 Illustrating how risks can be evaluated: This paper describes how risks can be evaluated in a privacy-preserving data visualization based on a systematic understanding of the different attack scenarios [1].

2.2 Anonymization Methods

The anonymization methods used in the context of visualization fall broadly into two categories, which are methods based on: i) data uncertainty and ii) visual uncertainty (Figure 2.2). Introducing uncertainty in the data space involves use the of the anonymization methods (Chapter 1) for making sure either a certain number of records are indistinguishable, and the distribution of attributes is such that sensitive information cannot be derived from them. Besides the traditional metrics of k-anonymity, l-diversity, t-closeness, and differential privacy, we also find examples in the literature where novel metrics are proposed. For example, Okansen et al., using a dataset of users' cycling work-outs [95], focus on three methods, namely privacy-preserving heat map with user diversity (ppDIV), privacy-preserving kernel density estimation (ppKDE) and privacy-preserving user count calculation (ppUCC). Their goal is to prevent the disclosure of user identity. Data-based clustering algorithms [96, 97, 98] and those based on differential privacy [99] are also used for preventing identity and attribute disclosures.

In visualization, at least some information about the data is typically available, like labels and value range on axes, and the minimum and maximum boundaries of each cluster. The notion of a totally 'blind' attack, without any knowledge about the data, may not be applicable to privacy-preserving visualization. To guard against this kind of inference based, researchers had proposed the idea of developing anonymization metrics in the screen-space, as opposed to the data space, based on visual uncertainty. Visual uncertainty [100] entails uncertainty owing to the visual mapping between data points and pixel coordinates. For example, a clustered scatter plot or a parallel coordinates (Figure 2.2b) that guarantees a minimum level of privacy, can be developed by combining pixel binning with the conventional anonymization methods like k-anonymity or l-diversity [101, 102].

Visual uncertainty has important connotations for how the intended privacy level of a visualization can be breached via different attack scenarios. As shown in Figure 2.3, the cluster ranges naturally hide record locations within a cluster and cluster overlaps can also hide where a record within a cluster ends up, across the axes, in a parallel coordinates plot. An attack usually consists of a series of progressive actions, building on incrementally acquired knowledge. An attacker may start with little knowledge, and by making observations from the information conveyed in visualization, such as a clustered parallel coordinates or a scatter plot, the attacker may try to identify a particular record within that cluster.

From that, the attacker gradually identifies more information about the record by moving from one axis to another or works out information about other records in the same cluster, as shown in the illustrations involving cluster overlaps cluster splits, and cluster range in Figure 2.3. Regardless of how complex an attack is, it can be decomposed into a set of basic attacking actions and disclosure risks. Causes and effects of visual uncertainty (in the form of cluster overlaps, splits and ranges) can protect against disclosure risks and computing the amount of uncertainty [101, 102] and can also provide an estimate to data owners and custodians of the degree of risk involved with different visualization configurations [1]. Other examples of visual uncertainty involve the use of record masking [85] or obfuscation for volume rendering [103].

2.3 Visualization Tasks and Techniques

Visualization has a key role to play in all aspects of privacy in the data ecosystem for both data owners and data consumers. With our dual focus on visualization-specific contributions and application-specific research involving privacy-preserving data visualization, we are able to cover a breadth of work that can inform both visualization researchers and practitioners. In this section, we describe the surveyed

papers based on the following categories (Figure 2.1): i) the high-level visualization tasks relevant to privacy-preservation, and ii) visualization techniques used to address those tasks. In this section, we describe the privacy-preserving data visualization tasks and techniques that we collected from our survey. Five high-level visualization tasks emerged in our collection, and we describe them along with the corresponding visualization techniques.

2.3.1 Hide data

Hiding data was the most common in our collection, with a coverage of more than 50% of the papers we surveyed. This task was employed for both spatial data and non-spatial data. In rare cases, we find the use of machine learning models for minimizing the exposure of sensitive information using a cloud-based architecture [104]. For scientific data, Chou et al. [103] proposed an obfuscation technique for scientific visualizations in order to maintain the privacy of the user. This block-based volume data transformation algorithm obfuscates volume data and delegates the task of rendering the volume data to a remote server, thus preserving the privacy of the scientific visualization. The images show the difference between normal rendering and the proposed privacy-aware volume rendering. This paper also demonstrated the development of a transfer function adjustment so that the transfer to the remote server for volume rendering is also privacy preserving.

For spatial data, the primary goal is to hide the exact coordinates of people's location [105]. To that end, Andrienko [106] presented a visual analytics model which can analyze the episodic digital traces/locations of a person over a long period of time and detect places of significant interest like home, work, social activity place etc. But this model also preserves the privacy of the person being analyzed. Geographical maps are used to represent neighborhoods instead of individual data points. It also uses a semantic map to display information about different places derived from the

data of a certain city. Two-dimensional time histograms are also used to analyze the usage of different location clusters in a certain city over a certain period of time. Ljubic et al. [107] use geographical heatmaps to present the distribution of influenza in a certain area. This helps in finding the affected area in a certain geographical region, which may be helpful to healthcare officials. A privacy leakage in these geographical heatmaps may allow the identification of certain patients, leading to identity disclosure.

For temporal data, visualization is often used to encode the outcomes of an anonymization method (e.g., k-anonymity, l-diversity, t-closeness, differential privacy), leveraging clustering in the data space [97, 98, 108] for visualizing event sequences.

For non-spatial data, visual uncertainty is added to a conventional technique like scatter plot or parallel coordinates as an additional defense mechanism [101, 109, 102]. Examples of visual uncertainty include loss of precision of a data point, where an attacker is unable to tell apart lines in parallel coordinates or dots in a scatter plot due to visual confusion, or the degree of granularity of records in a cluster, where an attacker is not able to exactly point to record locations within a cluster. Understandably, visual uncertainty can reduce the risks of both identity and attribute disclosure by manipulating clustering algorithm parameters.

2.3.2 Evaluate risk

Evaluating risk was the second most common task in our collection, with a coverage of about 30% of the papers we surveyed, mostly focused on the data owner. Disclosure risks are affected by how much an adversary knows about the data. Two kinds of re-identification scenarios are possible [78]: a) prosecutor re-identification scenario, where an intruder (e.g., a prosecutor) knows that a particular individual (e.g., a defendant) exists in an anonymized database and b) the journalist re-identification

scenario, where an adversary tries to randomly re-identify an individual based on the distribution of certain quasi-identifiers, demographic attributes, or even sensitive attributes. Researchers have recently proposed visual uncertainty-based risk quantification. Researchers in application domains like healthcare [50] discuss how privacy-preserving data sharing risks can be mitigated in a non-interactive privacy scenario, by restricting the queries that can be used for exploring the data. These concepts can also be applied in the case of interactive visualization, where different visualization configurations are evaluated carefully for risk factors before making them publicly accessible. Data owners thus need to rigorously identify risks before releasing the data. Kao et al. [87] present a novel visualization interface named ODD visualizer which will help in open data de-identification, i.e., if there is any privacy leakage in the dataset. It uses heat maps to display k-anonymity and l-diversity distributions. This is similar to the approaches of Castellani et al. [39], who propose a visualization-based data profiler for understanding potential vulnerabilities in openly available city data, and Deeb-Swihart et al., where they evaluate strategies to help law enforcement officials combat human trafficking while ensuring privacy protection [110]. Recently, Dasgupta et al. [1] proposed a suite of metrics using which data owners can estimate the probability of disclosure risks of different configurations of clustered scatter plots and parallel coordinates. The risk quantification model addresses both re-identification scenarios and quantifies the number of guesses an attacker had to make before knowing the precise value of an attribute or the location of a record within a cluster (Figure 2.3). Assessing these risks can help data owners decide on an appropriate level of privacy they are comfortable with, before releasing the visualization for public access. Another example of such a task includes the analysis of privacy preservation with human trajectory data [111]. Wang et al. conducted experiments to understand how a user can analyze movement behaviors using trajectories and how they can locate specific positions on these trajectories.

They observed that trajectory analysis is more accurate and even less time-consuming while using Positions of Interest (POI) than road networks or histogram but locating positions on a trajectory is almost the same in POI and Road network methods. This paper also comments that the capability of these features in trajectory analysis and privacy exposure may differ for various trajectories, based on the area covered. Thus the combination of multiple features may generate new knowledge, but it also increases privacy risk.

In one of the few examples focusing on evaluating privacy risks for a data subject, Takano et al. proposed a visualization system [112] for making users aware of how different entities for website tracking can potentially compromise user identity without their knowledge. In another such example, Muchagata et al. [113] presented a text-based interface in a mobile application that will help patients and healthcare professionals to monitor health data. The most important feature of this visualization, named Adaptive Graphical Visualization Interface (AGVI), is the interface is user-adaptive, i.e., it changes according to the user's needs. This paper observes that adaptive visualization techniques can influence the users' perspective on the security and privacy of a mobile application, but the roles of the user (patient or healthcare professional) and their goals (searching for medications or analyzing patients' tests) can influence this perspective. This is the only example where we found that an interface is tested with respect to multiple roles, and design considerations are presented from both a data subject and a data consumer's perspective.

2.3.3 Understand policy

Understanding Policy was the third most common task in our collection, with a coverage of about 25% of the papers we surveyed targeting both data subjects and data owners as users. Bahrini et al. [92] discuss how a mobile application

can help users to understand which user information is accessible by the granted permissions. This interactive visualization will help the users make an informed decision about whether to install a certain application or not. The authors claim that the results of their evaluation state that by promoting user awareness regarding permissions required by mobile applications (Android), users pay more attention to these permissions. The paper also tested system usability using error bars for different versions of the application and concluded that the version with a more detailed description/flow of permissions has greater usability. Dhotre et al. [94] implemented a method to perform a semi-automatic analysis of the privacy policies of certain websites and generate visualization in order to help the user understand the policies better. This visualization interface, consisting of pie charts, helps the user understand the use of different Personally Identifiable Information (PII) by the website, according to their privacy policies. The interface also summarizes certain sections, like the use of cookies and information sharing policies, and help the users to understand them better. The Privacy Policy Elucidator Tool (PPET) collects the privacy policies from different websites, parses them, classifies them using machine learning techniques like Naïve Bayes classifier, and uses the extracted paragraph and summary for the visualization. It also evaluates the trustworthiness of the website and displays the same through a donut visualization. Ghazinour et al. [114] present a visualization model which will help the data owners understand the privacy policy of a website and help the policy officers to better understand the designed policies. The Privacy Policy Visualization Model (PPVM) involves the use of relationship diagrams to help in the following tasks: understand privacy policies of these websites when using the name and email address of individuals to send notifications regarding new services, not collecting data of anyone under a certain age limit, disclose user information pursuant to lawful requests etc. The model suggests highlighting the purpose(P), granularity(G), visibility(V), retention(R), and constraint(C) of the privacy policies in this relationship diagrams.

Becker et al. [93] reflects whether using visualizations to communicate privacy and security measures have positive effects on trust. Infographics are used to depict certain privacy concepts like SSL encryption and AES encryption and study the improvement on privacy and trust. The study concluded that though these descriptive images have a positive effect on the trust in the provider, there was no significant improvement regarding data security and privacy in comparison to the text-based privacy policy.

2.3.4 Evaluate trade-offs

The task of evaluating trade-offs, performed mainly by data owners or custodians, had a coverage of about 18% of the papers we surveyed. Wang et al. [115] developed a combination of tree-based and matrix-based visualization techniques for helping data consumers dynamically understand the effect of privacy parameters on the difference between the original data and the processed data.They propose the construction of a Privacy Exposure Risk Tree for interactively controlling how hierarchical attributes are organized and selecting parameter values of a privacy model based on differential privacy. A matrix-based view is then used to observe the change in two-dimensional distributions of different combinations of selected attributes. At the end of this process, they can also export an anonymized dataset. Xiao et al. [88] presents a visualization tool named VISEE which will help to maintain the balance between high application utility and less privacy leakage in the case of sharing of sensor data. Accelerometer data collected from different mobile devices have been used as an example. The visualization focuses on representing the degree of mutual information between different pairs of variables. Parallel coordinates, feature grid diagrams, and ranking charts help select the appropriate combination of features and sampling rates, thus making a good decision on the trade-off between utility and privacy. For data subjects, Wang et al. proposed an interactive visualization tool for users who can share their personality portraits by tuning the privacy settings, visualized in the form

of linked bar charts [116]. Ragan et al. [85] presents an interactive interface where the user starts with fully masked de-identified data and later clicks to open when more information is required for making better decisions. This is a system that reduces privacy risk through on-demand incremental information disclosure. Box plots have been used to analyze the test results in different masking levels like full, moderate, low, and masked.

2.3.5 Compare algorithms

The task of comparing algorithms had a coverage of about 18% of the papers we surveyed, focused mainly on data owners to understand how different algorithms have an effect on privacy or re-identification risks. A significant challenge in incorporating multiple models is comparing the effectiveness of different anonymization schemes as privacy requirements can drastically change across datasets and user backgrounds. To address this problem, Wang et al. developed a tool called GraphProtector [117] that guides users based on the transformation steps in a privacy-preservation pipeline. Using interactive visualization in the form of a graph, users can manipulate sensitive and non-sensitive nodes and their connections and observe the structural changes to the graph that interferes with utility. Ultimately, they can make better decisions about which algorithm is appropriate for their data and privacy goals.

Kung et al. [118] use Discriminant Component Analysis (DCA), a supervised version of Principal Component Analysis (PCA) for the visualization because DCA can support data of high compression (small dimensionality), and the recoverability can be controlled. This paper has also compared the results of different clustering methods using multidimensional projections using which users can compare and effectiveness of this approach.

2.4 Critical Reflection on the Design Space

The goal of a conventional visualization or visual analytics technique is to facilitate the generation of insights from data. While the definition of insights itself has been debated by several researchers [119, 120], there is no denying the fact that visualization processes maximize the amount of information that can be encoded in and decoded from a visual representation. This is in contrast to the goal of any privacy-preserving data visualization technique, where the goal is to restrict data consumers from accessing sensitive information or help data owners understand the trade-offs and policies governing such restrictions. In this section, we aim to study how this contrast is reflected in the design choices. To this end, we refer to the literature on the ranking of channels [121, 122] and analyze the role of high-accuracy channels (e.g., position) and low-accuracy channels (e.g., area) for privacy preservation purposes. We include techniques from our collection and augment that collection with techniques that use either class of these channels. We first discuss a classification scheme (Figure 2.4) and organize our analysis around three themes: i) transformation of high-accuracy channels, ii) vulnerability of low-accuracy channels, and iii) the relative utility of these channels when a transformation is applied for privacy-preservation purposes.

2.4.1 Classification scheme

Privacy-preserving data visualization techniques use a transformation of the channels that would be otherwise used for visualizing the de-anonymized data. As part of our classification scheme (Figure 2.4), we group the techniques based on the **original channel** that is used for visualizing the raw data and for each of them, identify the **privacy-preserving channel**.

Paper	Original Channel	Visualization Technique	Low-Level Task	Vulnerability	Privacy-Preserving Channel	Modified Vis Task	Risk Source
DK11		Parallel Coordinates	Identify	Disclosure- both	Area	Summarize	Interaction
AAF16		Geographical Map	Locate	Identity disclosure	Density	Distribution	Interaction
Wang2017		Scatterplot	Identify	Attribute disclosure	Area	Distribution	Interaction
Kung2017	Position	Multidimensional projection	Identify	Identity disclosure	Containment	Group	Knowledge
Dasg14		Pixel-based	Detect trends	Attribute disclosure	Color	Detect trends	Interaction
Wang2017		Scatterplot	Distribution	Attribute disclosure	Area	Distribution	Knowledge
Mazzia2012		Multidimensional projection	Group	Identity disclosure	Containment	Same	Knowledge
DMK14	Height	Bar Chart	Compare	Attribute disclosure	Same	Compare	Distribution
DMK14	Area	Treemap	Compare	Attribute disclosure	Same	Compare	Distribution
CY16		Volume rendering	Detect shape	Attribute disclosure	Same	Detect shape	Knowledge
Dasg14	Shape	Glyph	Detect patterns	Attribute disclosure	Same	Same	Knowledge
Xiao2018		Scatterplot	Distribution	Attribute disclosure	Same	Same	Distribution

Figure 2.4 Dissecting the design space of privacy-preserving visualization: in terms of the transformation of the original channel (used for encoding the raw data) to a privacy-preserving channel. In particular, we point to the vulnerability of the high-accuracy channels like *position* and also highlight the counter-intuitive fact that even low-accuracy channels like *area* and *shape* can be exploited by attackers.

We use the task taxonomy proposed by Brehmer and Munzner [123] to distinguish between the high-level **privacy-preserving task** (i.e., *why* a task is performed) and the **low-level visualization task** (i.e., *how* a task is performed).

The main reason for a privacy focused transformation (e.g., a scatter plot transformed to a clustered scatter plot) is to prevent the original tasks from being performed owing to their vulnerability. Therefore, we also look at the **modified visualization task**, and introspect on the relative difference in utility between the original and the anonymized visualization. Finally, we also reflect on what possible **risks** could be associated with the anonymized visualization. Such risks can stem from the interactivity of a visualization, where additional context, or description is provided or from the background knowledge of an attacker.

2.4.2 Vulnerability of high-accuracy channels

In geographical maps and in multidimensional visualization techniques like scatter plots and parallel coordinates, the position is the primary encoding channel. Assuming that individuals are represented using these visualizations, a high-accuracy channel like position can help identify individuals and thereby leading to a privacy risk of identity disclosure. Privacy-preserving parallel coordinates and scatter plots have been proposed by generalization through k-anonymity [101], where records

are visualized as clusters. When the position visual variable provides the primary encoding, then we can exploit the difference in resolution between the screen space and data space to inherently lose information through binning, etc. This, when used as a parameter for controlling a privacy-preserving algorithm, can produce visualizations with both high privacy and utility. However, it has been shown that cluster-based k-anonymous parallel coordinates and scatter plots have certain vulnerabilities from record linkage and attribute linkage [102].

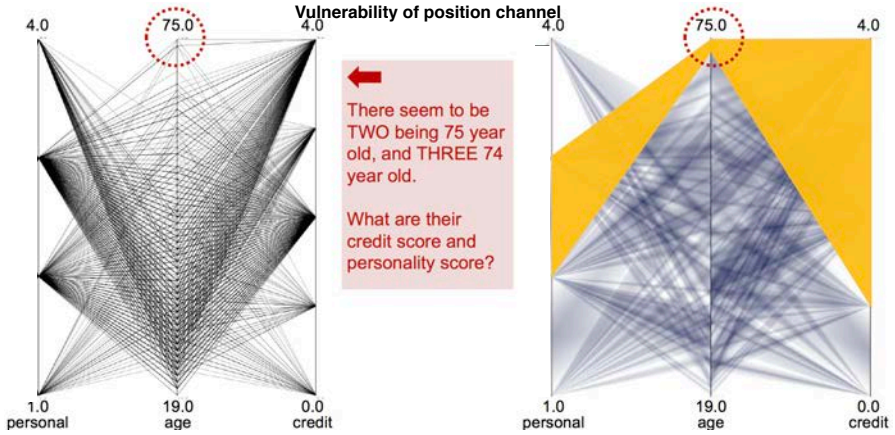


Figure 2.5 Illustrating vulnerability: In a position-based encoding, where clustering can help transform a position-based encoding to an area-based encoding and protect against sensitive queries.

An example of such vulnerability is shown in Figure 2.5. In this case, the edges of clusters represent real data points. If an attacker is aware about, say, the age of a person, as shown in the figure, and the pixel coordinate of that data point coincides with a cluster border, then the location of the record is revealed. On the other hand, if the pixel coordinate is a non-edge point within a cluster, that provides higher privacy. With respect to attribute linkage, one can geometrically derive the number of possible cluster configurations given different values of k and use that for guessing the linkage between adjacent attributes. Reordering and brushing can enable an attacker to choose a different adjacency configuration of quasi-identifiers and browse through a

subset of records. Dasgupta et al.[102] have proposed different screen-space metrics that aim to constrain such interactions based on the privacy risks.

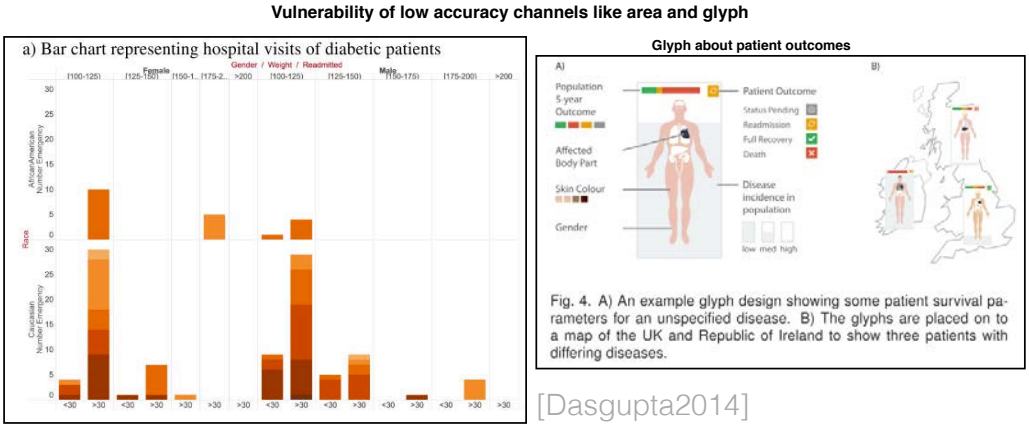
Transformation of the position channel to a density-based representation in geographical maps [106] is also common, where users can gauge the distribution instead of locating individuals. Such manipulation of pixels is also possible with non-spatial pixel-based visualization techniques, where value of an attribute is mapped to colors according to a chosen color scale [124]. But in the case of interactive pixel-based visualization [125], each pixel can be an entry point to an individual's data point, and malicious users can use a number of educated guesses to know the value of an attribute. Pixel-based representations can also become vulnerable when linked with other contextualizing representations.

Other approaches towards the transformation of the position channel include the use of containment metaphor in the case of multidimensional projections [118] and converting raw scatter plot representation to a representation of distributions [115]. While such transformations guarantee a minimum level of privacy, they are also vulnerable to interaction, especially drill-down operations, which should be adaptively restricted based on the associated risks.

2.4.3 Vulnerability of low-accuracy channels

Low-accuracy channels like area, density, shape etc., which generally represent aggregated data, can be intuitively thought of as being inherently privacy-preserving. In this case, one is unable to observe the exact value of an attribute or locate a record precisely. Yet, as demonstrated in earlier work [125], such an assumption is not valid in many real-world use cases.

As shown in the bar chart (Figure 2.6), some patterns stand out, like the correlation between the high re-admission rate and the number of emergency visits for male and female African Americans aged 50 to 60. There is only one category with



[Dasgupta2014]

Figure 2.6 Illustrating vulnerability in bar charts and glyphs: Despite aggregation and use of low-accuracy channels, information can be recovered using the data distribution or background knowledge.

non-zero frequency in re-admission greater than 30, and these are Caucasian males aged 40 to 50. This implies that with knowledge of quasi-identifiers such as race and age, deducing the diabetic condition would not be hard. Similarly, glyphs [126] can also be thought of harmless from a privacy-preservation perspective, nonetheless, as shown in the glyph in Figure 2.6, more information can be potentially determined about the patients based on the background knowledge of the attacker. Glyphs are popular visual representations in the healthcare domain because of the intuitive nature of the representation. In contrast, such information, when integrated with openly available attributes, patient identity can be at risk: using small DNA sequences from the Y chromosome, researchers at MIT were able to extract the genealogical information (surname, relatives) and religious background of fifty people from the 1000 Genomes Project [127]. The same rationale applies to the use of shapes in the case of volume rendering [103]. In summary, low-accuracy channels do not guarantee the preservation of privacy and appropriate risks should be assessed in the context of the externally available information about the individuals who are represented.

2.5 Gaps and Research Opportunities

Based on our survey, we present an analysis of the key gaps and research opportunities thereof. We organize this section based on research themes, each of which addresses the following key questions motivated by the well-known Helimeijer catechism [128]:

- What are the limitations of the current practices of privacy-preserving data visualization?
- Why is it important to address those limitations?
- How does a research approach or contribution look like, for addressing these gaps?
- Who will be the beneficiaries of the proposed research direction: data consumers or owners?

We believe these questions will help us understand both the significance of the research problem and the potential impact of the visualization-specific solutions. We sort the following research themes based on the authors' subjective understanding of the connection between related visualization research and the suggested directions: ones where there are immediate connections are presented first. This is, however, not a commentary on the importance or impact of the suggested research.

2.5.1 Uncertainty visualization and privacy

The lack of empirical evaluation of the effect of anonymized visualization on users' perception of privacy is a key gap in the literature. With the exception of a few [85, 129, 130], we did not find any other examples where controlled studies have been conducted to investigate how well the theoretical guarantees of privacy hold good in practice. Such studies will help data owners and custodians understand the following: how easy or difficult is it for people to breach privacy for a single dataset, how well users can leverage their background knowledge to breach privacy, and what other additional context can either be suppressed or controlled to add uncertainty or confusion in the minds of an attacker. In recent years, the broadly

defined research area of uncertainty communication has made a lot of progress [131]. As mentioned before, there is an inherent link between uncertainty and privacy: many anonymization methods can be treated as uncertainty quantification mechanisms and the added uncertainty due to visual mapping has already been termed as visual uncertainty. We need to conduct controlled studies with raw data and visualization with uncertainty encoding and measure the ability of users in terms of time and cognitive effort, to recover the identity of individuals or the values of sensitive attributes by overcoming uncertainty. It would be worthwhile to use Bayesian approaches for modeling how people's background knowledge and prior beliefs can lead to disclosure risks even in the presence of uncertainty in the visualization.

An application of quantification of visual uncertainty (i.e., the uncertainty resulting from the visualization process) is that different views of the data can be calibrated by their degree of vulnerability, in terms of disclosure risks, and interaction constraints can be enforced so that users are only able to access views that guarantee a minimal level of privacy. For multiple coordinated views, this means that details-on-demand [132] can be constrained based on privacy parameters in addition to the users' goals and needs.

2.5.2 Dynamic visualization of risks for privacy stakeholders

As pointed out recently by a study [51], there is a high degree of vulnerability of anonymized datasets, especially which contain demographic attributes, even after applying the state-of-the-art privacy-preservation techniques. With the proliferation of IoT-based devices and the evolving concept of smart homes [133], such vulnerability will need to be continuously evaluated by both data subjects and technology developers. This is a key gap in the literature, where privacy is considered only at the time of the release of a dataset, and data custodians do not have the tools to re-evaluate risks in the face of newly released datasets or other attack scenarios.

This gap makes most of the open data repositories vulnerable to privacy breach, even though personally identifiable information is not present in those datasets. With respect to visualization research, we found very few papers [129, 85] focusing on this aspect of privacy. There is a fertile ground for visualization research that aims at communicating vulnerabilities in open data and privacy-utility trade-offs to all stakeholders.

Visualization-based interfaces can play a key role in helping data owners, subjects, custodians, and consumers dynamically evaluate the disclosure risks of shared data. For data owners or custodians, visual interfaces [134, 39] can help communicate privacy risks by suggesting non-obvious, probabilistic linkages [135], let them dynamically evaluate the trade-offs among data utility and privacy risks [136] by visualizing privacy outcomes from new and evolving metrics [137], and make more confident decisions regarding data sharing [138].

2.5.3 Privacy-aware citizen science

Developing smart cities with the help of data collected about citizens' mobility patterns, preferences, habits etc., is a potential which has attracted the attention of governments across the world. But this also means that data about people's location and movement are more vulnerable than ever before. The New York Times report [139], which we pointed to earlier, and shows about the ease with which people's location can be known, is alarming. While this cannot be solved simply by applying computational techniques, this issue is symptomatic of the opaque ways in which urban data is collected and administered. A study had previously demonstrated how urban mobility data collected by analyzing New York City taxi trips can compromise the identity of individuals [140]. This is a research gap relevant to both data owners and data subjects, as it is the individual's data that is collected and analyzed in this case. While we have encountered several

papers [141, 106] focusing on privacy issues of spatial data, such research needs to be integrated more deeply with the research involving privacy-preserving urban data collection [142, 143] and decision-making. Research grounded in behavioral sciences has recently demonstrated the benefits of using visualization-based interfaces for granting citizens the transparency to directly administer and understand the implications of data sharing [144]. Visualization techniques need to be further developed and explored for more inclusive and transparent citizen science, where third-party interference can be minimized, and citizens can more proactively exercise their right to privacy.

2.5.4 Ethical data visualization through privacy by design

Researchers in computing and data-driven technologies are becoming more cognizant of the moral obligations and ethical implications of research [70]. Automated analysis, machine learning, and provenance should be controlled, and it should allow those impacted by the decisions to appeal their decisions or seek better outcomes. We have certain ethical obligations as visualization designers, as we generally have complete access to data and the freedom to portray insights derived about people. When we are presenting data to the public, as visualization designers, an abiding principle should be to protect the privacy of the people whose data we have collected and visualized, even if at the cost of communicating our key findings. Both data and data visualization are not ethically neutral activities, and thus there is an obligation to be ethical while representing data [145]. Integrating principles of "privacy by design" [146] in visualization interfaces will be a key research opportunity to this end. Moreover, as natural language interfaces begin to be integrated with visualization techniques [147] and visualization techniques begin to be augmented with text-based facts [148] care should be taken that the computationally generated facts are privacy-preserving as well.

2.5.5 Interpretable privacy policy-making

In the face of new legislations like the GDPR and questionable practices by online companies about privacy policy communication, we foresee a significant amount of research effort being dedicated towards interpretable policy-making: where both data subjects and data owners can better understand privacy parameters before implementing policies and data consumers can overcome the barriers of intended [47] or unintended [46] obfuscation for better understanding policy implications. In our collection, we encountered several papers [94, 93, 114, 86] dedicated towards studying this problem. But most of this research is concentrated on application-specific domains. Greater collaborative efforts across domain experts and visualization designers can significantly improve the quality of the visualization techniques we encountered. In many of these cases, we found data-flow diagrams, relationship diagrams or infographics being used as a means of communicating policies. Except for Dhotre et al. [94], we sensed a lack of quality in the visual communication of information extracted from policy text. We believe that recent advances in text visualization and topic modeling [149] can have a significant effect on improving visualization techniques for communicating privacy parameters and their dependencies, as extracted from policy descriptions, and make that information accessible and actionable, especially for data subjects, who might not have appropriate levels of data literacy to comprehend the privacy risks and policies.

2.5.6 Privacy-preserving and inclusive visualization

Many recent studies have shown that it is the poor and marginalized section of society, who are in the greatest danger of violation of their privacy rights [150, 151]. In our collection, we found research focusing on law enforcement agencies that collect data about potential human trafficking involving vulnerable people [110]. Care needs to be taken to preserve the privacy of these data subjects, who are vulnerable and

do not have access to services otherwise guaranteed in urban areas. With the proliferation of smartphones and wearable technology [152], visualization techniques can be used to collect data from people who need assistance, legal, social, or otherwise. Inspiration can be drawn from a recent study on visualization perception in Rural America [153], and visualization can be used as a privacy-preserving data collection medium, where people can "see" themselves as part of a larger societal structure and can also get assurance about their privacy not being violated. Visualization designers and researchers have a unique opportunity to be inclusive of the marginalized and underrepresented population while, at the same time, respecting the ethics of preserving privacy.

2.6 Conclusion

We live in times of constant threat to individual privacy, where all of us are mere data points as part of some data-driven digital commodity. There are many risks to such massive collection and aggregation of data, where data can be de-anonymized, and individuals can be re-identified without their consent for malicious purposes. In this survey, we have reflected on the challenges and opportunities that we face in the visualization community, with respect to the larger socio-technical challenge of privacy-preservation. One of the key opportunities for the field of privacy-preserving data visualization is to develop novel solutions with data subjects as the stakeholder, many of whom are often at the receiving end of uninterpretable privacy policies or are exposed to greater privacy risk, since they come from vulnerable sections of the society. We believe there is scope for immediate impact with all the research directions outlined above. They will help us progress along the path of resolution of the ongoing and ever-increasing dichotomy between individual privacy and data-driven consumerism.

CHAPTER 3

DISCOVERY OF VULNERABLE DATASETS

3.1 Problem Characterization

Open data portals democratize access to hitherto proprietary data, thus, encouraging participation from both data custodians and data subjects. Data custodians, like governments, can use this to make or improve policy decisions, while *data subjects*, like citizens, can use this to understand their participation in society. Portals like NYC Open Data [10], Kansas City Open Data [11], and City of Dallas Open Data [12] host datasets across different domains like healthcare, economy, infrastructure, and others.

The risk of disclosure is exacerbated with the rise of these open data portals that collect citizens' data and publish de-identified versions of these data, which can be further used for research purposes. Though these portals improve the accessibility of government data, thus, promoting transparency in governance, it also leads to an important question: what if datasets within the open data ecosystem are linked even without any other sensitive information from private datasets? Rocher et al. showed that 99% of Americans can be re-identified even from heavily anonymized datasets using a combination of demographic attributes like date of birth, gender, zip code etc. [51]. Another study re-identified individuals from the de-identified medical records of only 10% of the population, released by the Australian Government Department of Health [13]. Lavrenovs and Podins showed how the privacy of passengers could be violated through the public transportation open data, released by the city municipal of Riga, Latvia [14]. The issue of re-identification or disclosure of sensitive information has also been addressed by other researchers, thus, pointing to the need to investigate the privacy issues of open datasets [154, 155, 156]. Recent

studies have also shown that the risk of re-identification may vary over time and is dependent on the number of datasets available at the time of analysis [157]. But, the general practice of "release-and-forget" followed by data custodians and owners (henceforth referred to as data defenders) aggravates this problem of re-identification since generally, the open datasets are not reviewed for their risk posture after their release [158, 159].

This calls for a comprehensive study into the different possible disclosure vulnerabilities present in the open data ecosystem. Thus, in order to understand the risk of re-identification or disclosure of sensitive information by joining datasets from open data portals, we performed a red-teaming activity where we donned the hat of an ethical hacker to understand the attacker's perspective and identify the vulnerable entry points into the open data ecosystem. Then we report some of the vulnerabilities observed during the red-teaming exercise into the open data ecosystem and present some other possible vulnerabilities that may arise in the future. This helped in the development of a risk inspection workflow, named PRIVEE, along with the tasks required to replicate the different attack strategies. These tasks are then realized in a web-based visualization interface with the specific goal of implementing a defender-in-the-loop analytical framework that can be privy to the disclosure risks or possibility of inadvertent leakage of sensitive information whenever new datasets are released. In this chapter, we first discuss the red teaming exercise through the vulnerabilities detected during that time. This is followed by the discussion around the development of a dataset of highly susceptible datasets, which was later used to develop the PRIVEE workflow.

3.2 Red-team Exercise

A red-team exercise can be generally defined as a structured process to better understand the capabilities and vulnerabilities of a system by viewing the problems

through the lenses of an adversary[160]. In the context of security and privacy, red-team exercises follow the cyber kill chain by playing the role of an ethical hacker and emulating the possible attack scenarios[161]. With the help of researchers in data privacy and urban informatics, we performed a red-teaming exercise by inspecting the open datasets for vulnerabilities. We engaged in a cold-start exploration process, followed by a more focused exploitation of datasets with privacy-relevant attributes, for developing a shared mental model of the problems related to the vulnerabilities. The intuition here was that the datasets with universally known quasi-identifiers, like age, race, gender, etc., can lead to the disclosure of sensitive information when joined with other such datasets.

3.2.1 Attack through vulnerable entry points

Red-team exercises generally follow the cyber kill chain. It starts with the initial reconnaissance step, where attackers try to find vulnerable entry points into any target system. Moreover, attackers used *quasi-identifiers* [48] like age, race, gender, and location to breach privacy by linking multiple datasets [49]. Inspired by this, we bootstrapped our red-teaming activity by searching for datasets with these known quasi-identifiers. During our initial exploration, analysis of these datasets led to interesting observations where some of the datasets have a highly skewed distribution of records across different categories of the quasi-identifiers.

For example, the dataset *Whole Person Care Demographics 2* [162] from the *County of San Mateo Datahub* portal [163] had only one record for a 26-year-old female of the Hawaiian race. This can lead to identity disclosure and leak of sensitive information when joined with other datasets. Another dataset *Demographics for Public Health, Policy, and Planning* [164], from the same data portal, had only seven records for age 19. However, out of these seven people, only one person was male. This individual can be identified since other identifying attributes like race, language,

and city were also present. This may also lead to attribute disclosure if other similar datasets are exploited.

The dataset *Overdose Information Network Data CY January 2018 - Current Monthly County State Police* [165] from the *Pennsylvania Open Data* [166] portal had only few records for the race American Indian/Alaskan Native. But, out of these few people, only one was of Hispanic ethnicity. This dataset also contained location attributes, thus, making it easy to pinpoint an individual with these attributes. When joined with other publicly available information, this information may lead to identity and attribute disclosure for this individual.

Thus, datasets with vulnerable entry points can be exploited to reveal sensitive information about a human data subjects. The presence of such datasets in the open data ecosystem can be considered a warning sign that calls for the development of a method that acts as the trusted informer for data custodians and informs them of potential disclosures in a proactive manner.

3.2.2 Attack exploiting dataset joins

The previous attack scenario established that vulnerabilities exist in individual record-level datasets. This leads to an essential question of whether these datasets can be actually joined with other open datasets to expose sensitive information. Join is a fundamental operation that connects two or more datasets, and joinability is the measure to determine if two datasets are linkable by any number of join keys [167, 168]. When these *join keys coincide with protected attributes* like age, race, location, etc., the outcome of the join can potentially reveal sensitive information about an individual or even disclose the individual's identity.

As a next step in the red-teaming exercise, we randomly selected vulnerable pairs of datasets from multiple open data portals [10, 11, 12] and analyzed them for *joinability risks*, in terms of what kind of sensitive information may be leaked by

these joins. Several iterations of the selection of joinable pairs and join keys led to the discovery of disclosure between the datasets *Juvenile Arrests* and *Adult Arrests* from the *Fort Lauderdale Police Open Data Portal* [169]. We observed that two individuals, aged 15 and 21, mentioned separately in these datasets, were involved in the same incident of larceny on 20$^{\text{th}}$ March 2018, at the Coral Ridge Country Club Estate, Fort Lauderdale. This can be an example of identity disclosure by joining two open datasets. Further investigation revealed other examples where two individuals, aged 17 and 21, mentioned separately in these datasets, were involved in the same incident of motor vehicle theft on 8$^{\text{th}}$ of July, 2018. The presence of linking attributes like *case id* between datasets *Adult Arrests* and *Citations* helped to reveal an incident where a 28-year-old black male who was arrested for larceny on 26$^{\text{th}}$ September 2021 at NW 10$^{\text{th}}$ Ave, Fort Lauderdale, was also cited for disobeying stop/yield sign and driving while license is suspended at NW 9$^{\text{th}}$ Street, just around 3 miles away from the arrest location. A similar incident was also observed while joining datasets *Citations* and *Juvenile Arrests* on the linking attribute *case id*. In this incident, a 17-year-old white male was first charged with disobeying a red light. He was later arrested for possession of cannabis over 20 grams on 4$^{\text{th}}$ August, 2015, both at N Federal Hwy, Fort Lauderdale.

We repeated this exercise and found other examples where dataset joins ultimately led to disclosures. In another example, two datasets, namely *Electronic Police Report* 2016 and *Electronic Police Report* 2015 from *New Orleans Open Data* portal [170], were joined on quasi-identifiers like *location*, *victim age*, *offender age*, *victim race*, *victim gender*, and *offender gender*. On inspection of the joined records, we observed that a 22-year-old black male was charged with attempted robbery with a gun against a 27-year-old white male at 6XX Tchoupitoulas St on 13$^{\text{th}}$ July 2015 at 01:00 hrs and again on 30$^{\text{th}}$ April 2016 at 03:00 hrs with attempted simple robbery. This is an example of identity disclosure even when *masking techniques* are used

on the address. Another observation from these joined records revealed an incident where a runaway female juvenile of age 17 was reported at 85XX Dinkins St on 26th February 2015, and the same incident was closed through a supplemental report one and half years later on 7th December 2016. Incidents like these may be rare; hence, identifying the individuals from these records may not be difficult.

We also observed such examples across other open data portals. Datasets *APD Arrests Dataset by Neighborhood*, and *APD Field Interview Cards Dataset by Neighborhood* from the *Albany Police Department* [171] were joined on the attributes *age*, *race*, *sex*, and *neighborhoodxy*. We observed that a 28-year old white male was interviewed by the police in the Washington Park neighborhood at 09:08 hrs on 1st December, 2020 and was later arrested for trespassing on enclosed property at 12:32 hrs. This leads to attribute disclosure for the individual arrested as the arrest details are revealed. Joining other datasets like *APD Arrests Dataset by Patrol Zone* and *APD Field Interview Cards Dataset by Neighborhood* from the same data portal revealed similar incidents where a 26-year old black female was interviewed at 11:23 hrs on 11th December, 2020 and was later arrested at 21:25 hrs for "assault with intent to cause physical injury". In another example, joining datasets *APD Field Interview Cards Dataset by Neighborhood* and *APD Traffic Citations by Neighborhood* on a broader set of attributes like *age*, *sex*, *neighborhoodxy* and *date* led to another interesting observation related to a police incident. We observed that a 23-year old male was stopped for a field interview on 23rd January, 2021 at 2:45 am. Since field interviews are usual routine stop and search activities by the police, this may seem a regular incident. But the other dataset informed that an individual of the same age and gender received a citation on the same date and at the exact location at 2:48 am, just 3 minutes after the incident from the first dataset. Since both these records seem to belong to the same person, this is a possible identity disclosure, and it was

discovered using a combination of date and quasi-identifiers like location coordinates, age, and gender.

3.2.3 Attack through transitive dataset join

Inspired by these examples and the concept of transitive dependency in databases [172], we explored the concept that two datasets, which have no shared attributes between them, can still be joined if they have shared attributes with a third dataset. We experimented with different permutations of dataset joins in order to find an example of transitive disclosure. Though we did not find any examples of transitive disclosure at this stage, this can be an interesting field of research that can further strengthen the inspection of disclosure risk in open datasets.

3.3 Development of Vulnerable Datasets

Open data portals contain a multitude of datasets on varying topics like economics, health, and others. However, they may not be relevant in information disclosure about human activity. On top of that, the problems discussed in the previous section press for an urgent need for a smaller subset of open datasets focused on disclosure risks. Hence, we curated a seed set of datasets that contains a subset of the open datasets, which may be more susceptible to vulnerabilities related to disclosure. In this section, we discuss developing this dataset and the learning outcomes.

3.3.1 Data collection

Many open data portals are developed using frameworks/APIs like Socrata API [40], CKAN API [173], DKAN API [174], etc. We selected the Socrata API as our source for the open datasets. Though other APIs could have served a similar purpose, we planned to start with Socrata and develop a generalizable approach that can help integrate the other publicly available APIs.

First, we queried the list of all available data portals through Socrata Discovery API. From each of these data portals, we queried the metadata for all the data items available within them. Data items include datasets, maps, data dictionaries, etc. We filtered these results and created a list of $39,507$ datasets.

3.3.2 Data analysis

Manually analyzing all these datasets would be a difficult task for any analyst. Thus, we developed a semi-automated program that filters datasets if they have some combinations of the known quasi-identifiers. Initially, we started with a list of the known quasi-identifiers like *age*, *sex*, *race*, and *age group*, to name a few, and the program selected the datasets with these attributes. After evaluating the attribute space of the selected datasets, we subsequently updated this list to include more such quasi-identifiers. This helped us to select a broader set of datasets that may be susceptible to disclosure risk through these quasi-identifiers.

Multiple iterations of this process led to the development of a set of about 5404 datasets with some combination of the quasi-identifiers. We also manually verified the unselected datasets to check if we had missed any such vulnerable datasets.

3.3.3 Data curation

After reducing the set of candidate datasets, the next step was to determine if these datasets relate to human objects and activity. Hence, we started manually curating the metadata file to understand what each dataset pertains to. For each of the datasets, we opened them in their respective data portals and analyzed them to understand if they were related to human data subjects or not. We observed many such datasets with location attributes (like zip code, address, etc.). Nonetheless, many of them do not necessarily relate to human beings, like *datasets for street lamps*, *building details*, etc. We dropped those datasets since they are irrelevant in the context of the privacy of the data subjects.

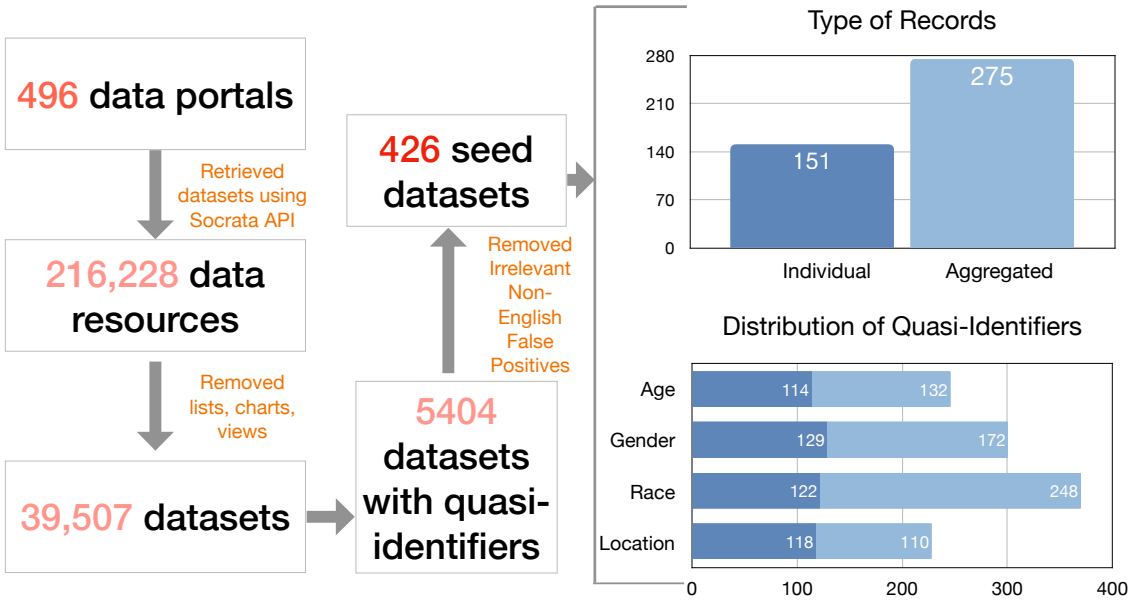


Figure 3.1 Dataset development: The dataset development process starts with over 216,000 data resources from 496 data portals. After a few filtering steps, it consists of 426 highly susceptible datasets with different levels of granularities and distribution of quasi-identifiers.

Removing these datasets related to non-human objects, we curated a seed set of 426 datasets of varying granularity. 151 of these datasets were individual record-level (e.g., records of people committing crimes) while the rest 275 datasets were aggregated record-level (e.g., college records) datasets (Figure 3.1). We understand that a dataset collection like this should be continuously updated. However, we need more infrastructure to set up a method to fetch and update this collection regularly. Thus, we also plan to release a metadata file for this dataset collection as a contribution to this dissertation.

3.4 Discussion

Identifying disclosures using traditional search options in open data portals is challenging. Moreover, data custodians might need more information than shown in the search results to find disclosures. Thus, this context demands a visual analytic

system specifically targeted toward disclosure evaluation and other privacy pitfalls. Our workflow PRIVEE, described in the later chapters, can be considered as an initial attempt toward this purpose. The visual analytic design space explored in PRIVEE helps establish a streamlined workflow responsive to the data custodian's inputs yet distilling the results effectively.

However, this system can have users other than a data custodian. During the development of the workflow, we realized that a data subject could also be interested in discovering if their data can be compromised by exploiting these privacy pitfalls. Our work can address the data subjects' perspective too. But an approach leveraging an individual user's attribute values may be more efficient in this context. Hence, we envision that future design solutions in this space will be more geared toward the data subjects' perspective. This can be incredibly beneficial in encouraging data activism by citizens [175, 176, 177].

Another attack scenario we envisaged during the red teaming exercise is the disclosure of sensitive information through the transitive join of open datasets. We are still leading a separate effort toward quantifying the transitive disclosure risk. The primary challenges in this effort are the presence of limited examples yet a high number of possible combinations to explore. This may serve as an important field of research since disclosures like this are difficult to detect by data custodians, yet they can have a massive impact on the privacy of the data subjects. We hope researchers look into different visual analytic solutions to address this attack scenario.

3.5 Conclusion

Open datasets are essential in improving government transparency and empowering citizens with access to hitherto proprietary data. We discuss some of the privacy pitfalls of open datasets with real-world examples we observed during an ethical hacking exercise. This highlights the importance of addressing the privacy pitfalls

on an urgent basis. Towards that end, we develop a collection of highly susceptible datasets that help effectively emulate the strategies developed during the exercise and identify disclosures. We also envision exploring possible disclosure risks beyond joinable pairs and improving the web-based interface's data processing capabilities in collaboration with big data experts. We believe this dataset and the vulnerabilities we observed will be used to develop more effective solutions and help data defenders safeguard the interests of the open data ecosystem.

CHAPTER 4

PRIVEE: DISCLOSURE INSPECTION WORKFLOW

4.1 Introduction

Accessibility of open data portals (e.g., NYC open data [38]) is like a double-edged sword. On the one hand, they make institutions and organizations accountable by providing public access to proprietary information. On the flip side, inadvertent data leaks could compromise the privacy of data subjects. Recent research has shown how the lack of checks and balances in the conventional release-and-forget model [51] makes it surprisingly easy to breach privacy. An underlying reason for such a high privacy risk is the joinability of multiple open data sets that contain information about people. However, data owners and custodians (hereafter referred to as defenders) lack effective ways in which joinability risks can be summarized and communicated at the time of data set release or whenever a vulnerability is detected online.

As discussed earlier, several recent examples of privacy breach scenarios emphasize the urgent need to address this problem. The Australian Department of Health released *de-identified* medical records for 2.9 million patients (10% of the population), but researchers were able to re-identify the patients and their doctors using other open demographic information [13]. Passengers' private information might be disclosed through the public transportation open data released by the city municipal of Riga, Latvia [14]. Researchers were also able to re-identify the details for 91% of all the taxis in NYC using an anonymized open taxi dataset and an external dataset [140].

Complete automation of the risk evaluation process is not feasible due to several reasons, like the presence of noisy metadata and the requirement for human expertise. Noisy metadata hinders the automatic profiling of these datasets. The various

definitions and temporal nature of privacy risks, owing to the intermittent release of new datasets, point to the necessity for a human-in-the-loop approach, where defenders can configure and update risk computation techniques based on evolving compliance needs.

To address this critical need, we conducted a red-teaming exercise in the form of a design study with urban informatics and data privacy researchers to develop a **proactive risk inspector** that is privy to the sensitive information that can be leaked before and after dataset release in urban, open data portals. PRIVEE, the visual analytic workflow resulting from this design study process, acts as a data-driven risk confidante and informer for the defender in the analysis loop. PRIVEE emulates potential attack scenarios and enables defenders to triage risky dataset combinations and ultimately diagnose the severity of disclosed information through dataset joins. A defender can thus proactively check for risks while releasing a dataset or depend on PRIVEE to be alerted when new vulnerabilities emerge owing to newly available, joinable data.

As the first contribution of this design study, we characterize the problem of disclosure evaluation and develop a set of visual analytic tasks that can be executed in a workflow to detect, calibrate, and inform data defenders about disclosure risks (Section 4.2). These tasks, developed in collaboration with privacy experts, emerged when we analyzed the problem through the lens of an adversary and developed several attack scenarios during the red teaming exercises. We observed that it is possible to breach the privacy of open datasets using these scenarios, thus corroborating the findings of NYC taxi data in a larger scope where we can find information about data subjects [140]. As our second contribution, we designed the visualizations required for implementing the PRIVEE workflow and let defenders explore and interpret risks at the **metadata level**, triage vulnerable dataset groups and corresponding high-risk **joinable dataset pairs**, and ultimately reason about the

severity of the information disclosed at a **record-level** The design of these techniques is rooted in the idea of automation with transparent explanations which are responsive to user-controlled risk configurations (Sections 4.4, 4.5, 4.6). Finally, we present an interactive interface to help data defenders execute the workflow and demonstrate its effectiveness in the end-to-end diagnosis of disclosure (Section 4.7) through two case studies with domain experts.

4.2 PRIVEE Workflow and Tasks Characterization

The results from the red-teaming exercise confirmed our intuition that datasets with quasi-identifiers, when linked together, can potentially divulge sensitive information. Analyzing the functional requirements, we, together with our collaborators, concluded that totally automating the risk evaluation process is infeasible as human intervention is necessary at multiple stages of risk definition, interpretation, and subsequent exploration of the dataset combinations at high risk. To formulate a solution, we collaboratively developed PRIVEE, a visual risk inspection workflow in which defenders can proactively engage to stay one step ahead of the attackers (Figure 4.1).

PRIVEE is motivated by protecting the most vulnerable data sets against data join attacks. The workflow serves the dual purpose of: i) observing the open datasets to detect potential privacy vulnerabilities and ii) being a trusted informer for the data defenders that can visually explain and communicate disclosure risks while encouraging a deeper exploration of the attack and defense strategies. Automating the analysis of the disclosures directly at the record level can be an alternative, but this may lead to a seemingly infinite number of combinations to explore. Our streamlined workflow, developed from the experience gained during this design study process, will help the data defenders focus on a set of highly vulnerable datasets, thus reducing the number of combinations to be explored. In this section, we first describe the inputs

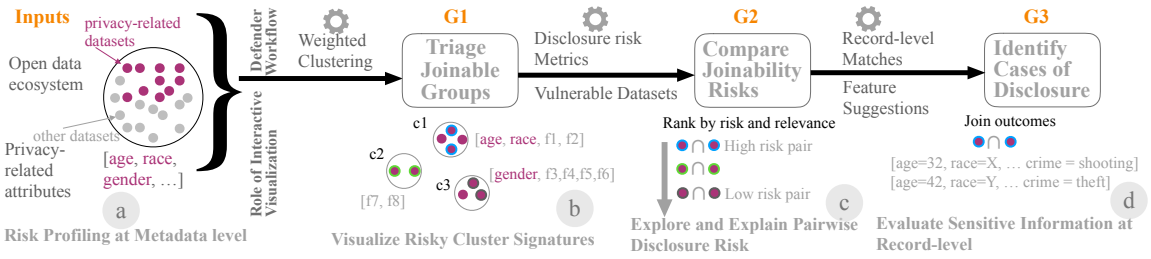


Figure 4.1 PRIVEE is an end-to-end risk inspection workflow for open datasets: It informs the defender in the analytical loop about potential disclosure risks in the presence of joinable datasets. Interactive visualization plays a crucial role in bootstrapping the risk inspection process via risk profiling, triaging and explaining risk signatures, and ultimately detecting instances of true disclosure at a record level. Colored borders track datasets across the goals.

and then define the high-level goals of the PRIVEE workflow in order to map them to the corresponding visual analytic tasks ultimately realized in a web-based interface.

4.2.1 Inputs to the workflow

We initiate our defense strategy on the seed set of privacy-related datasets, which are about people as the data subjects, that we collected during the red teaming activity. While collecting these datasets, we followed the universally accepted common quasi-identifiers like age, race, gender, etc., with the notion that an open data ecosystem should, at a minimum, protect against attacks using these well-known quasi-identifiers.

After carefully curating the metadata from the seed datasets, we observed that there is no standard nomenclature for the attributes across the different data portals. This lack of standardization established the importance of creating a metadata dictionary, starting with the well-known set of quasi-identifiers like age, race, gender, and location, and focusing on the well-known quasi-identifiers while providing defenders the guidance and flexibility to define other privacy-related attributes. These attributes and the datasets selected based on their metadata serve as the inputs to the PRIVEE workflow (Figure 4.1a).

4.2.2 Triage joinable groups (G1)

Candidate datasets for inspection selected from the initial input can be of the order of tens or hundreds. Finding all possible combinations of dataset joins among them is computationally expensive. Moreover, the large set of join outcomes will not lend well to human interpretation of risk. Also, during the red-teaming exercise, we observed that the risky datasets could also be construed from the datasets with vulnerable data distributions. Therefore, the next tasks in the defender's workflow are to focus on groups of datasets that can be joined and then triage those groups based on risk indicators:

T1: *Explore cluster signatures:* As shown in Figure 4.1b, this task lets defenders explore cluster signatures in terms of presence (clusters c1, c3) or absence (cluster c2) of the privacy-related attributes and their overall semantics. Involving the defender ensures that their inputs influence the algorithms used for grouping, using weighted clustering. They can thus control the triaging process by judging the groups' risks and privacy relevance. This task ultimately helps them select clusters of interest for further inspection of joinability risks.

T2: *Find vulnerable datasets based on data distributions:* The red-teaming exercise highlighted the presence of disclosure risk in datasets with a highly skewed records distribution across different categories of the quasi-identifiers. This task helps to distinguish between the most vulnerable and other datasets by inspecting a high likelihood of finding unique records for given quasi-identifiers.

4.2.3 Compare joinability risks (G2)

Once a cluster of datasets is prioritized for inspection as part of G1, defenders would like to compare joinable pairs of datasets in this group that may potentially disclose sensitive information. To achieve this goal, we use disclosure risk metrics to automatically suggest risky pairs based on their feature profiles and then visualize

those suggestions so defenders can interpret the metrics. The following task achieves this:

T3: *Explore and Explain Disclosure Risks:* This task focuses on pairs of datasets that can be ranked using multiple disclosure risk metrics. Within those rankings, we want to use visual cues that directly explain: which features are responsible for high risk, the differences between high and low-risk pairs, and if other features should augment the defender's definition of privacy relevance.

4.2.4 Identify cases of disclosure (G3)

Once dataset pairs are selected as part of G2, defenders would like to understand the severity of the join outcomes. Fully automating this process may lead to many scenarios where the disclosures are less concerning and do not warrant any significant change in the defense strategies. To provide more control to defenders in their diagnosis of cases of actual disclosure, the tasks required to accomplish this goal are:

T4: *Detect matching records across data sets:* Matching records are the records present in both datasets in a pair. The main objective of this task is to detect lower frequencies of matching records, which may lead to the disclosure of sensitive information about an individual or disclose their identity.

T5: *Augmenting the risky feature set with suggestions:* One way of discovering disclosures is finding attributes that have the same values for all the records of the joined datasets. For example, joining two hospital datasets may reveal that all the patients common in both the hospitals are treated for cancer, leading to attribute disclosure for these patients. In this task, we suggest a set of attributes that may be highly related to the joining attributes, thus helping the users augment the feature set for the dataset join.

Figure 4.2 Interface Design: The design of PRIVEE comprises rich interaction among filters and multiple views: (a) Filter area helps select datasets based on metadata like tags, data granularity, and privacy-related attributes; (b) Projection View lets the defenders compare the signatures of different joinable groups of datasets and evaluate vulnerable data distributions; (c) Risk View helps compare the risk for dataset pairs and select the high-risk pairs; (d) Disclosure Evaluation View helps to analyze the matching records for potential disclosures.

4.3 Design Overview

The design of PRIVEE is motivated by the need for a transparent explanation and evaluation of the risk inspection process. We implemented a web-based interface that enables data defenders to iterate between multiple entry points, evaluate the reasons for the dataset joinability and analyze disclosure risks for different combinations of datasets and attributes. In this section, we provide an overview of the design requirements for realizing the aforementioned visual analytic goals and tasks.

Risk Profiling at metadata level: PRIVEE helps to analyze the datasets' risk profiles through a filter bar, located conveniently at the top of the

interface (Figure 4.2a), which contains a search option for the different tags and options to select the data portals and the dataset granularity. During the initial page load, this filter bar is positioned at the center of the page in order to avoid overwhelming the user with the search results. Defenders can select any combination of the tags from the tags search option, which is enriched with a modified bar chart showing the frequency distribution of the tags. Though the tags are sorted in descending order, the grey bar in the background (achieved by tweaking a linear-gradient bar) provides an idea of the frequency distribution of these tags among all the collected datasets. Privacy-related attributes can also be selected using filters.

Triaging joinable groups: In order to fulfill G1, PRIVEE employs a set of visualizations to help the data defenders triage the joinable groups from the datasets selected using their metadata. This includes a projection plot, a word cloud, and a bar chart depicting the attributes' frequency, as illustrated in Figure 4.2b. This combination of visualizations is repeated for the different groups of joinable datasets. Though PRIVEE automates the grouping of the datasets, these visualizations provide the data defender a transparent method to understand the group signatures and update the groups based on their domain knowledge and definition of privacy relevance.

Finding vulnerable datasets: PRIVEE helps the data defenders select vulnerable datasets by showing a distribution of the values of the privacy-related attributes through a combination of histograms (for numerical attributes) and bar charts (for categorical attributes), as shown in Figure 4.2b. This combination is repeated for each dataset, ranked according to their degree of vulnerability. It is also responsive to the privacy-related attributes selected through the filter area. The vulnerable categories for these attributes and their labels are shown in bright red to help defenders efficiently select vulnerable datasets.

Comparing Joinability Risk: PRIVEE automatically computes the possible pairs from the datasets selected from either the Projection View or the Vulnerable Datasets View and ranks them according to their joinability risk. The visual cues, shown in Figure 4.2c, help the data defender compare different datasets and select the high-risk pairs on a priority basis. Overall information about the risk score distribution allows flexible selection of dataset pairs of varying risk.

Identifying disclosures: The disclosure of sensitive information can depend on multiple factors, subject to evaluation by the data defender. In this *Disclosure Evaluation View*, as shown in Figure 4.2d, PRIVEE lets the data defender analyze the matching records generated for a specific dataset pair and a join key selected from the Risk Assessment View. PRIVEE also suggests other features to help the defenders select a better join key, helping them understand the relationship between different attributes and possible disclosures.

4.4 Triage Joinable Groups (G1)

Data defenders need to analyze the degree of joinability between datasets. Hence, the design requirements for addressing tasks T1 and T2 are to develop human-in-the-loop clustering methods responsive to multiple definitions of privacy relevance, along with transparency in analyzing cluster signatures. This enables defenders to develop a mental model of the context and the degree of the potential vulnerability of subsequent joins. In this section, we discuss the analytical methods and visualizations to find and triage the joinable groups.

4.4.1 Weighted clustering for finding joinable datasets

Converting Data Attributes to Word Embeddings: The joinability of two datasets is a function of *shared attributes*. Thus, the datasets with similar attributes should be more joinable. Attribute names in open datasets are often noisy and inconsistent, making it computationally difficult to perform a binary search for the

presence or absence of certain attributes. We focus on the idea that similar attribute names can capture the semantic similarity among multiple datasets that might have a similar context. We use a word-embedding approach that simultaneously satisfies the need to capture datasets' joinability and their semantic similarity. *Word embeddings* can be defined as real-valued, fixed-length, dense, and distributed representations that can capture the lexical semantics of words [178, 179]. Thus, we converted the data attributes into their corresponding word embedding form using Python's spaCy library [180] and created a vector representation for the attribute space of each dataset. The vectors with a smaller distance between themselves signify datasets with similar attributes, therefore more joinable.

Adding Weights for Privacy-related attributes: At this stage, all the data attributes have equal importance in the vector representation of a dataset; hence, datasets with attributes like *version*, *version number*, etc. may be marked similar to each other. But these attributes may not have much significance in the context of privacy. Therefore, we decided to add weights to some of the privacy-related attributes identified from the seed dataset corpus. Attributes like *age*, *race*, *gender* and *age at arrest* were selected, and adding more weights to these attributes signifies that datasets having these attributes may be marked as more joinable. Any disclosure using these datasets can be considered a high risk, which will help further triage the datasets.

Cosine similarity is widely used to measure the similarity between words and documents [181, 182]. However, word embeddings are mere representations of the words, and multiplying them with numeric weights would not increase the cosine similarity between two datasets. Thus, we introduced a *weight vector* where we assign a weight if the privacy-related attributes selected by the data defender are present in the dataset. If a data defender selects the privacy-related attributes [*age, gender, race*], then the corresponding weight vector for a dataset with only the age and gender

attributes would be $[x, 0, x]$, where x represents the weight assigned to the privacy-related attributes. We concatenate these weight vectors with the corresponding word embedding vectors to get the final vector representation of each dataset.

Projecting the datasets and finding Clusters: Each dataset is now represented by a vector with more than 300 elements/dimensions, and comparing these datasets using a two-dimensional (2-D) or three-dimensional (3-D) plot would be challenging if all the dimensions were used. Hence, we used the t-SNE dimensionality reduction algorithm to reduce these into two-dimensional vectors [183]. A 2-D projection of the datasets might not readily reveal dataset groupings. Thus, we experimented with clustering algorithms like KMeans [184], DBSCAN [185, 186], Birch [187], and OPTICS [188, 189]. After a careful analysis of the clusters' quality and the cluster density scores, we selected the DBSCAN algorithm.

Evaluating the clusters: There can be multiple groups of similar/joinable datasets, which would lead to the creation of multiple clusters. A data defender may find it challenging to evaluate all of these clusters. Therefore, we have employed a few cluster evaluation techniques to triage these clusters **(T1)**.

One of such metrics is the *Calinski-Harabasz Index* which is defined as the ratio of the between-cluster dispersion and the inter-cluster dispersion, where dispersion means the sum squared distance between the samples and the barycenter [190]. A higher score signifies that the different clusters are far away, implying better cluster formation. We designed an experiment to evaluate the difference in the results from this metric along with other metrics like Silhouette Score [191] and Davies-Bouldin Index [192] and selected the Calinski-Harabasz Index since we observed that it could efficiently guide defenders in finding meaningful, joinable datasets.

Finding vulnerable data distributions: A particular cluster can have multiple datasets with vulnerable data distributions, leading to the disclosure of sensitive information when joined with other individual record-level datasets. Hence, we

found such data distributions and ranked these datasets according to their degree of vulnerability **(T2)**.

In order to evaluate the degree of vulnerability, we first analyzed all the datasets and created the record points for the privacy-related attributes present in them. Record points are the unique categories for a specific attribute, while *vulnerable record points* are those record points that have very few records for them, as shown in Table 4.1. These datasets are then sorted based on the number of such vulnerable record points present and the frequency of the most vulnerable record point. The intuition here is that a dataset with more vulnerable record points is more prone to disclosure risk using these privacy-related attributes.

Table 4.1 Sample Record Points

Record points	**Description**
[“age”, 11, 1]	For age=11, there is only 1 record
[“age”, 15, 5]	For age=15, there are 5 records
[“gender”, “F”, 2]	For gender=“F”, there are 2 records

4.4.2 Visualizing joinable group signatures

We designed the Projection View to provide an overview of the datasets and the joinable groups **(T1)** and perform an automatic evaluation of the vulnerable data distributions of the datasets in each joinable group **(T2)**. Data defenders can review the group signatures through the different components of the Projection View and update the parameters to see the details and the data distribution of the datasets that match their mental model of privacy relevance. The components of these views are described as follows:

Joinable groups: Given a set of datasets selected based on their metadata, defenders need to find groups of datasets that can be joined together. The analytical process is performed automatically by PRIVEE, leading to the formation

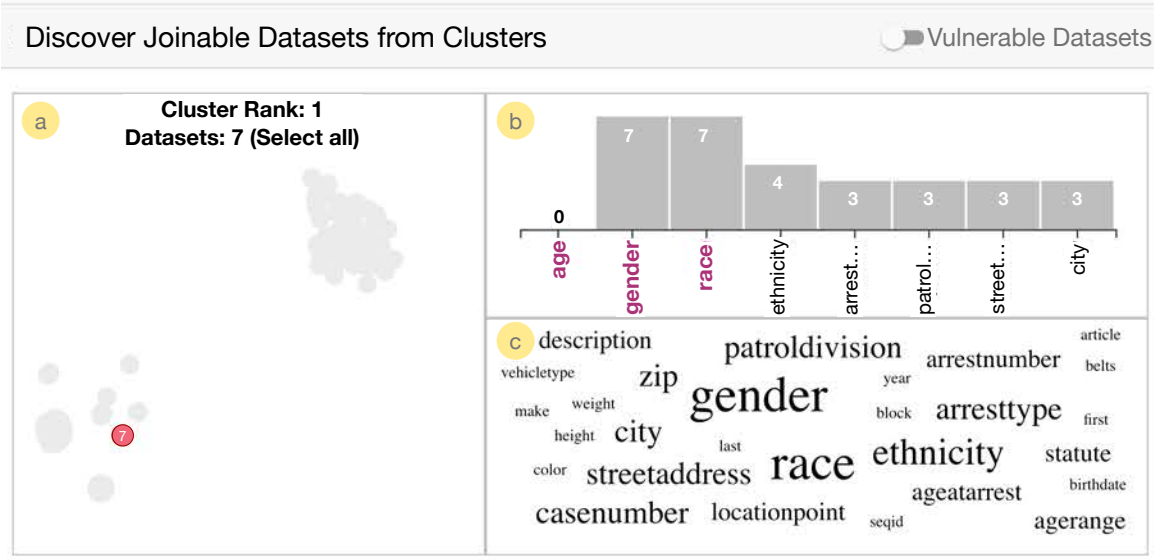


Figure 4.3 Projection View: A group of joinable datasets is represented in this view using (a) a projection plot. The (b) frequency distribution bar chart and (c) a word cloud for the attributes of a group of joinable datasets help in the transparent explanation of the group signatures.

of joinable clusters, which are represented using a multi-dimensional projection plot, as illustrated in Figure 4.3a. Here, a red dot represents an individual record-level dataset in a particular cluster, while the grey dots represent the datasets not in that cluster. During this design study, we realized that some of the datasets are highly joinable due to their similarity in the attribute space, which would cause overlapping of the dots in a cluster. Hence, the overlapping datasets are represented by a single dot with the number of overlapping datasets inscribed in it. For example, Figure 4.3a shows a cluster of seven highly similar datasets represented using a red dot. This view contains multiple projection plots, where each plot represents a group of joinable datasets. It helps the data defender quickly compare the different groups from a single view. The dual color encoding scheme (red-grey) helps visually differentiate between the datasets in a group and the other datasets. Initially, a scatterplot with different colors for the different clusters was also considered for this view. But it was realized that it is challenging to assign perceptually different colors to each cluster when the

number of clusters is large, due to the limits of perception. Hence, a multiple plot design approach was chosen with the two-color encoding scheme.

Transparent explanation of joinability and vulnerability: Understanding the cluster signatures is crucial in understanding the reason behind the genesis of a joinable group **(T1)** and the presence of data vulnerabilities **(T2)**. Since we have construed these dataset groups based on the similarity in their attribute space, it is essential to understand the frequency of the attributes present in these groups. As a result, bar charts become the natural choice for displaying the most frequent attributes in a group and their frequency, as illustrated in Figure 4.3b. These bar charts are sorted according to the attribute frequency, yet the frequencies of the privacy-related attributes are shown first. The vulnerable datasets are also represented using bar charts (for categorical attributes) / histograms (for numerical attributes) for each of the privacy-related attributes present in them. Bar charts can have the limitation of visual scalability where only a certain number of bars can be shown due to space constraints [193]. In order to overcome this limitation, we also introduce word clouds of the attributes, as shown in Figure 4.3c. All the attributes present in at least two datasets in a joinable group are represented in this word cloud, with the size channel representing their frequency.

The bar chart in Figure 4.3b explains the similarity of the datasets since all seven of these datasets have *gender* and *race* attributes, thus transparently explaining the group signatures. Besides overcoming the visual scalability limitation of the bar chart, the word cloud also helps the data defenders look for other attributes of interest that may have a lower frequency but have much larger relevance in the context of privacy. For example, attributes like *victim age* and *offender age* may not be significant for a general user; however, a data defender working with law enforcement may find them interesting since these attributes are used in police datasets. PRIVEE enables the data defender to update the default selection of the

privacy-related attributes, which triggers a re-rendering of the whole Projection View, thus automatically calculating new groups of joinable datasets with extra weightage to the newly added privacy-related attributes *victim age* and *offender age*. Together, these Projection View components enable human-in-the-loop dataset grouping that is adaptive to various definitions of privacy relevance by transparently displaying measures to evaluate cluster signatures.

4.5 Compare Joinability Risks (G2)

Dataset groups from the Projection View can lead to multiple pairwise combinations of datasets, where the data defenders need to analyze each pair for their joinability risk. Thus, the design requirement for addressing G2 is to facilitate efficient visual comparison of the risk profile of dataset pairs and guide defenders towards focusing on high-risk dataset pairs. In this section, we describe the metrics that can help a data defender quantify the risk of joinability between the candidate datasets and the subsequent use of visual cues to compare and prioritize the joinable pairs.

4.5.1 Metrics for joinability risk comparison

Multiple metrics that can help the data defenders compare the joinability risks between different dataset pairs were explored during the design study process. In this subsection, we define the mathematical formulas for the different metrics that highlighted the joinability risks better and were selected as part of the PRIVEE workflow.

Metric based on attribute profile: Shannon's entropy is a measure of the uncertainty of a random variable [194]. It has been widely used as a privacy metric [195, 196, 197, 198], as higher entropy signifies more unique values for that attribute, thus resulting in higher disclosure risk. Hence, we used this metric to help defenders find joinable attributes for a pair of datasets. For a pair of datasets (say A and B), we first calculated Shannon's entropy of each of their shared attributes

according to Equation (4.1) and kept their maximum as the entropy score for that attribute. The intuition here is that the attributes with higher entropy can be offered as suggestions to the defender for the join key.

$$H(X_J) = -\sum_{i=1}^{n} P(x_{J_i}) \ln P(x_{J_i}) \tag{4.1}$$

where X_J represents attribute X in dataset J ($J \in \{A, B\}$), $H(X_J)$ represents the entropy of an attribute present in dataset J while x_{J_i} represents each category of the attribute X_J in dataset J.

Metric based on dataset pairs in a join: Since the joinability of two datasets depends upon the number of shared features/attributes between them, the joinability risk score can be calculated as a function of the number of shared attributes and the number of privacy-related attributes between a pair of candidate datasets. The formulae for the *joinability risk score* can be defined as follows:

$$\text{risk} = \alpha * p + (c - p) \tag{4.2}$$

where α is the empirical risk ratio (a constant), p is the number of privacy-related attributes and c is the number of shared attributes.

The joinability risk score depends on the empirical risk ratio, and to determine its value, we designed an experiment to calculate the risk scores of all the possible combinations of joinable pairs from the seed datasets (${}^{426}C_2 = 90,525$ combinations). We observed that the value $\alpha = 50$ works well to separate the dataset pairs with privacy-related attributes and pairs without them; hence, the empirical risk ratio was fixed at the value of 50.

4.5.2 Visual risk assessment

PRIVEE uses multiple visual analytic components to encode the joinability risk metrics, and these components together form the Risk Assessment View. This

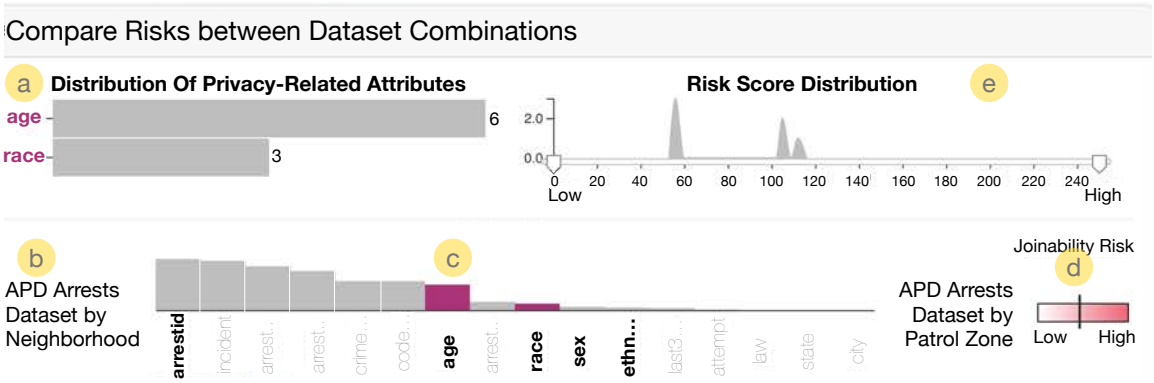


Figure 4.4 Risk Assessment View: (a) The distribution of privacy-related attributes can affect the joinability risks between (b) dataset pairs. Data defenders can compare the risk between these pairs by analyzing the (c) sorted bar chart showing the shared attributes and the joinability risk score represented by the (d) risk score bar. They can use the (e) risk score distribution histogram to focus on the dataset pair of their interest.

subsection describes how we map these metrics with the components of this view so that data defenders can proactively analyze the risk between the candidate datasets.

Comparing shared attributes set: The shared attributes' entropy metric encodes the attribute profile information, potentially highlighting if an attribute should be included in the join key. In the Risk Assessment View, these attributes and the entropy are represented using a descending *sorted bar chart* between the dataset names, as illustrated in Figure 4.4c. The horizontal position shows the different attributes, while the vertical position encodes the entropy of these attributes. The bars for the privacy-related attributes are colored in violet (plum kingdom), while the other bars were colored in grey, thus following the similar colorblind-safe two-color strategy used in the other views. During an initial design iteration, each shared attribute was represented using a small rectangular box, with each box containing the attribute name in it. However, we realized that this design leads to the loss of information about the difference in entropy between the different shared attributes. This led to the current design of the sorted bar charts where the data defender can

analyze the entropy, select any number of the shared attributes as the join key for the dataset pair and evaluate them for disclosures.

Comparing risks: Each dataset pair (Figure 4.4b) is represented with a combination of the following components: dataset names, shared attributes, and the joinability risk bar. These pairs are sorted according to the risk score. Thus, a top-ranked dataset pair would imply higher chances of joinability. In order to highlight the joinability risk score between the dataset pairs, the Risk Assessment View has a *joinability risk bar* for each dataset pair **(T3)**, as shown in Figure 4.4d. This bar is filled with a linear gradient between the grey and red colors, representing low-risk and high-risk dataset pairs. The exact risk score is highlighted using a black vertical bar. The choice of the colors, following the two-color scheme used across the different views in PRIVEE, helps express the joinability risk score on a scale of low to high scores. This view also shows an overview of the shared privacy-related attributes and the risk score distribution between the dataset pairs using a horizontal bar chart and a histogram (Figure 4.4a and Figure 4.4e). PRIVEE also automatically selects the joining attributes based on their entropy and privacy relevance, which the data defender can further augment.

4.6 Identifying Disclosures (G3)

The design requirement for addressing tasks T4 and T5 is to let the defenders *judge the degree of sensitive information* that can ultimately be disclosed through the joins. Since an apriori definition of risky features is insufficient, PRIVEE also suggests additional features to defenders for diagnosing sensitive matches. In this section, we first discuss the methods used for evaluating the disclosures, followed by the design of the visual cues that can help evaluate them.

4.6.1 Methods for disclosure evaluation

During the red-teaming exercise, we realized that the join key could vastly influence the disclosure of sensitive information. In this sub-section, we discuss two methods for disclosure evaluation:

Based on the low frequency of matching records: *Matching records* are the number of records present in the joined dataset. Thus, the presence of matching records can indicate the possible disclosures at the record level. However, the number of matching records may vary according to the choice of attributes in the join key and the type of records present in the datasets. For example, when joined on attributes x and y, dataset A and dataset B may have 200 matching records, but when joined on the attributes x, y, and z, they may have only 20 matching records. This implies that the attribute combination x, y, and z have a better chance of discovering an actual disclosure than the combination x and y. We have also observed that matching records may contain duplicates if the original datasets have duplicate or blank entries.

Based on the mutual information between the joining attributes: The selection of the joining attributes is an iterative process in PRIVEE. Mutual information measures the amount of information one random variable contains about another [199] and quantifies the mutual dependence of the two attributes of a dataset. Hence, we use normalized mutual information to suggest other features that defenders can use for detecting disclosures. PRIVEE automatically calculates the normalized mutual information between the joining attributes and the other attributes of the joined dataset. Next, it finds the top-5 attributes with the highest mutual information score and lets defenders consider those features for detecting matches **(T5)**.

4.6.2 Visual cues for evaluating disclosures

The design of the Disclosure Evaluation View follows Shneiderman's mantra [132], where PRIVEE first provides an overview of the matched records, then allows the

defender to explore them, and finally lets them view the record details on demand. Here we discuss the comparative visual cues [200] that aid in disclosure evaluation:

Exploration of matching records: Parallel Sets is a visualization method for the interactive exploration of categorical data, which shows the data frequencies instead of the individual data points [201]. PRIVEE shows the matching records using a modified parallel sets visualization, as illustrated in Figure 4.2d. Here, each attribute of the join key is represented using a stacked bar, where the height of the stacks represents the frequency of the different categories of that attribute. In the case of a numerical attribute, a histogram replaces the stacked bar and shows its data distribution. The numerical data is then divided into four equal bins to map them with the categories of the other join key attributes. The parallel sets for the privacy-related attributes are colored in violet, while that for the other attributes are colored in grey, following the similar color scheme used in the other views. The categories across the numerical and categorical attributes are connected using ribbons. Each ribbon represents the number of records in the joined dataset belonging to both categories. A simple click interaction on any of these ribbons opens a pop-up window showing the details of the records represented by the selected line.

This design helps detect both identity and attribute disclosures through the matching records **(T4)**. The thickness of the line may represent the identity disclosure, while the height of the stacked bar shows the attribute disclosure. For example, if there is only one record with a certain combination of all the join key attributes, this would be represented by a thin ribbon across the parallel sets visualization. This may potentially lead to identity disclosure if an individual is uniquely identified with this combination of the join key. Suppose if an attribute has only one category, then the corresponding stack height would cover all the height allocated to a certain attribute, revealing that all the individuals belonging to both the datasets have a particular feature and leading to attribute disclosure.

This Disclosure Evaluation View helps the data defenders ascertain the degree of the sensitive information disclosed by visualizing the overall relationship between the different attributes of the matching records yet retaining the granularity of the dataset at the record level.

Suggesting potential joining attributes: PRIVEE uses bar charts and histograms to encode the top-5 features with high mutual information with the join key attributes. These suggestions are positioned on the left and right-hand sides of the parallel sets, representing the feature suggestions from either of the datasets (Figure 4.2d). The privacy-related attributes are also highlighted in violet, while the others are colored in grey, following a color scheme similar to the interface's other views. Selecting any attributes from the feature suggestions would also update this visualization to include the newly selected attributes. These attributes can be used as suggestions for improving the initial set of joining attributes **(T5)**. The data distributions and the ranking of the attributes help boost defenders' understanding of the risky feature set that can be used as the join key.

4.7 Case Studies

4.7.1 PRIVEE as a risk confidante

In this subsection, we report a case study that our data privacy collaborator and co-author co-developed using the web interface of PRIVEE. He is a senior researcher with more than 15 years of experience in privacy-preserving data analysis and used PRIVEE as a privacy auditor. Specifically, he wanted to determine if there are any disclosure risks with the health-related datasets published in the open data portals and validate the role of PRIVEE as a risk confidante for data defenders.

Our collaborator selected the aggregated datasets in the interface PRIVEE along with the privacy-related attributes *age* and *race*; and then filtered them with the keyword "health" (see Figure 4.5a). He also enabled the Vulnerable Datasets switch to

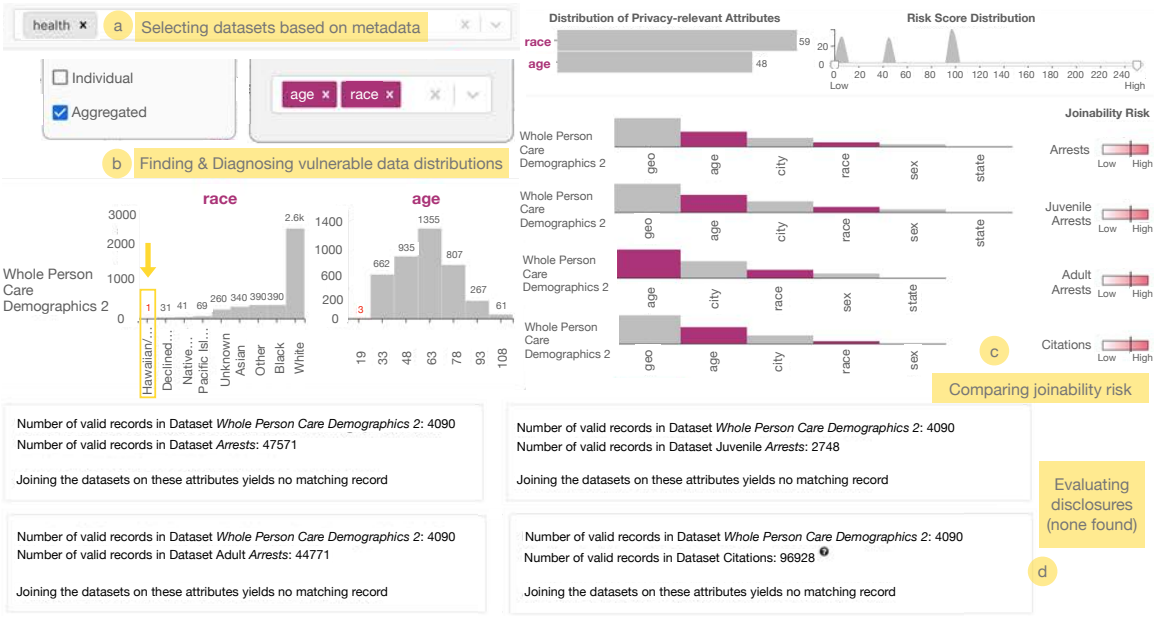


Figure 4.5 PRIVEE as a risk confidante for defenders: (a) Selecting datasets based on their metadata like the popular tag "health" and their granularity of records, (b) finding and diagnosing the vulnerable data distributions and observing that there is only 1 record for the race "Hawaiian", (c) comparing the joinability risk with the individual record-level datasets and (d) evaluating the disclosures with the top 4 individual-level datasets and observing that there is no disclosure.

check if there are any vulnerabilities in the data distributions of these datasets. At this point, our collaborator observed that the first few clusters do not have such vulnerable datasets. But the fourth cluster has the dataset *Whole Person Care Demographics 2* [162] from the open data portal of San Mateo county [163]. This dataset had only 1 record where the race was Hawaiian (Figure 4.5c) (**T2**). This was a significant cause of concern since if somebody knows a person in that county who identified as Hawaiian, then any dataset with a similar race category could potentially expose her health records. Thus, he started analyzing the risk of joining this dataset with all the individual-record level datasets available through PRIVEE, as shown in Figure 4.5c (**T3**). He decided to join these dataset pairs on the selected privacy-related attributes and the location attribute *geocodedcolumnn* since he wished to find datasets containing information relevant to this location. He observed that none of the top-4 dataset pairs yield any matching record when joined on these attributes (Figure 4.5d). Thus, our

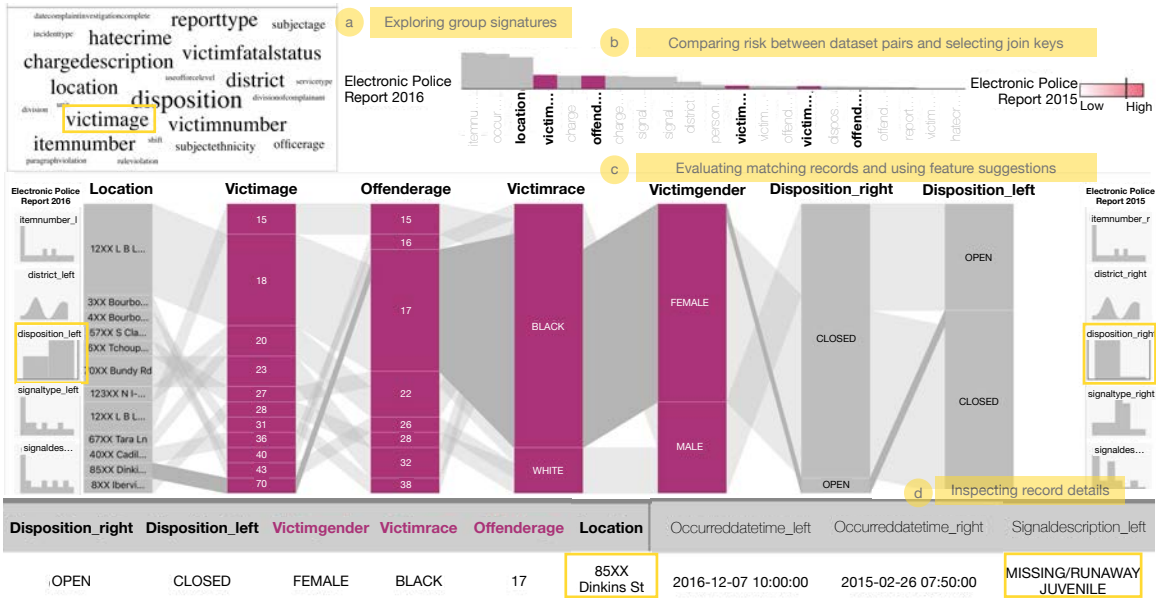


Figure 4.6 PRIVEE as a trusted informer for defenders: (a) Understanding group signatures and updating privacy-related attributes, (b) comparing the risk between dataset pairs, (c) evaluating the matching records using the feature suggestions shows that only one incident was open in 2015 but closed in 2016, (d) inspecting record details shows that a runaway juvenile can be identified despite the location being partially masked.

collaborator concluded that though this aggregated dataset has a meager count of a particular race, it does not lead to any disclosure (**T4**). He also analyzed a few other vulnerable datasets similarly but found no disclosures. Thus, PRIVEE acts as a risk confidante for the data defenders where they can analyze the disclosure risks for the vulnerable datasets in the presence of other open datasets. He also observed that he had not seen a tool with similar capabilities for interactive risk calibration and triage and commented: "*this is a great visual tool to explore privacy risks of open data, with the ability to visualize privacy risk across datasets in a dynamic manner*".

4.7.2 PRIVEE as a trusted informer

We report a case study that a researcher developed using the PRIVEE web interface. He is a senior researcher and university professor with over 25 years of experience in the fields of big data, cyber security, and scientific visualization. He focused on validating the role of PRIVEE as a trusted informer for the data defenders.

The researcher started by choosing the New Orleans Open Data portal [170] and observed 7 datasets on the Projection View, which were so similar in their attribute space that they were displayed using an overlapping circle with the number of datasets inscribed. Using the attribute distribution bar chart, he observed that none of the default privacy-related attributes (age, race, gender) were present in this group of datasets. However, on analyzing the word cloud, he made an interesting observation that attributes like *victim age* and *offender age* were present in these datasets, as shown in Figure 4.6a **(T1)**. Since, from his background knowledge, he knew that these attributes are generally present in police datasets, he updated the list of privacy-related attributes to select some of the similar attributes like *victim age*, *victim gender*, *victim race*, and *offender age*. As PRIVEE helps to triage the joinable groups of datasets based on the data defender's definition of privacy relevance, the Projection View was updated to reflect the change in privacy-related attributes.

He selected all these seven datasets in order to compare the joinability risks of the 21 possible pairwise combinations in the Risk Assessment View **(T3)**. Since he wanted to focus only on the high-risk pairs, he filtered out the low-risk pairs using the Risk Score Distribution histogram. Joining the first pair of datasets, the researcher observed that there are no matching records between them.

Next, he selected a pair of datasets, namely *Electronic Police Report* 2016 and *Electronic Police Report* 2015, but augmented the PRIVEE-suggested join key attributes and made the following selection: *location*, *victim age*, *offender age*, *victim race*, *victim gender*, *offender gender*, as illustrated in Figure 4.6b **(T3)**. He joined these datasets and observed 14 matching records in the Disclosure Evaluation View. He inspected further details about a certain record and observed that a 22-year-old black male was charged with attempted robbery with a gun against a 27-year-old white male at $6XX$ Tchoupitoulas St on 13^{th} July 2015 at 01 : 00 hrs and again on 30^{th} April 2016 at 03 : 00 hrs with attempted simple robbery **(T4)**. Next, from the

feature suggestions offered by PRIVEE **(T5)**, he selected the attribute *disposition*, which shows the status of a particular incident. He observed that only one record was open in 2015 but closed in 2016 (Figure 4.6c). On inspecting further details, as shown in Figure 4.6d, he found out that an incident of a runaway female juvenile of age 17 was reported at $85XX$ Dinkins St on 26^{th} February 2015, and the same incident was closed through a supplemental report one and half years later on 7^{th} December 2016 **(T4)**.

The researcher concluded that this is an example of identity disclosure where individuals were identified using PRIVEE even when the addresses were partially masked in de-identified datasets. He was also shown an earlier version of the PRIVEE interface during the case study. He commented that the new changes improved the rich functionalities of PRIVEE and added that this interface could help experienced data custodians analyze disclosure risks and potentially find examples of disclosures.

4.8 Discussion

When plugged into the open data stream, PRIVEE can act as both a risk profiler and a trusted informer that oversees risks while providing an appropriate level of control to defenders for integrating their domain knowledge using an end-to-end workflow. One of the lessons learned during this design study is that an interface helping defenders evaluate disclosures should enable seamless communication across sources and implications of risks while responding to the myriad definitions of privacy relevance. PRIVEE is bootstrapped by a default view that quickly adapts to the data defenders' inputs, allowing them to leverage appropriate levels of control while automating parts of the analysis process.

In its current implementation, one of the limitations of PRIVEE is scalability, concerning the number of records of each processed dataset and the size of the seed

input that is used for bootstrapping. We have limited the number of records to $100,000$ to avoid interaction latency.

There is also the need to incorporate greater automation in the selection of privacy-relevant, personal datasets without manual intervention. During this design study process, we learned that automation of this workflow is inherently challenging as privacy-relevance is subjective and open data are noisy; hence, training a model to mimic human judgment is difficult. Our approach of specifying a seed set outside the PRIVEE workflow is an important methodological choice allowing us to focus on the most vulnerable datasets and anticipated attack scenarios. Currently, PRIVEE only assesses joinability risk between pairs of datasets. It is certainly possible that there could be other scenarios like when multiple datasets are joined progressively, the risks propagate through the links. However, based on the feedback of our data privacy collaborator, we consider the risk scenarios handled in PRIVEE to be the necessary first steps toward assessing more complex combinations and variants of disclosure risks.

4.9 Conclusion

PRIVEE, the visual risk inspection workflow described in this design study, is a first step towards allowing data defenders both the control and efficiency needed to minimize disclosure risks from the joinability of open datasets. Through our case studies with data privacy experts, we demonstrated a key takeaway that the visualizations and interactions were effective in end-to-end exploration and diagnosis of the actual disclosure of sensitive information or identity of individuals. As an ongoing and future work, we will be exploring disclosure risks beyond joinable pairs. We will further augment our workflow with intelligent and scalable data processing capabilities in collaboration with big data experts. We also plan to conduct controlled

studies to evaluate the usability of PRIVEE and its components with real-world cyber defenders.

Acknowledgment

The work reported in this chapter was supported by the National Science Foundation (CNS-2027789) and the National Institutes of Health (R35GM134927). The content is solely the responsibility of the authors and does not necessarily represent the official views of the agencies funding the research.

CHAPTER 5

VALUE: UTILITY CALIBRATION WORKFLOW

5.1 Introduction

The linking of open datasets can create valuable insights for addressing specific problems. For instance, the records of two companies' customers can be combined to identify overlapping records and reveal customers who have patronized both companies. Similarly, the records of police arrests and court proceedings can be merged to extract more comprehensive information about individuals included in both datasets. The open data revolution, founded on the FAIR data principles, has increased the accessibility of such datasets [35]. This growing accessibility can enable researchers to discover new opportunities for joining open datasets to gain deeper insights. However, the open data ecosystem can be considered a forest of datasets, presenting a challenge in leveraging their value through dataset linking. Quantifying the value gained from joining these datasets and selecting dataset pairs with higher utility are complex tasks. Therefore, transparently evaluating the utility of various open dataset combinations has become critically important.

To overcome these challenges associated with joining open datasets, we develop a user-configurable utility metric that expresses the value of pairwise dataset joins based on these datasets' attributes and record space. This metric is then leveraged to develop the VALUE framework and a web-based interactive visual interface, enabling researchers to compare the utility of joinable open datasets and calibrate it. But manually performing pairwise joins and evaluating their utility can be cumbersome and time-consuming due to the sheer scale and complexity of the combinatorial explosion that arises when dealing with a large number of datasets. For example, with a group of 400 datasets, there can be up to $80,000$ potential pairwise combinations,

highlighting the need for automating the computing processes to evaluate the utility of these combinations efficiently. While the JOSIE algorithm uses a similar automated approach to identify joinable tables in large data lakes using set similarity techniques, relying solely on automation may overlook valuable insights that can be gained from the user's input and background knowledge, making a human-centric approach necessary [202]. Our approach enables interactive triaging of joinable dataset pairs by human stakeholders (e.g., social science researchers) leveraging the combination of a new utility metric with a visualization interface for distinguishing between the most and the least useful joinable pairs.

In this chapter, we first understand the different join scenarios through examples (Section 5.3). This understanding is then leveraged to contribute the utility metric that can triage the joinable and useful dataset pairs from a large group of datasets (Section 5.4). Next, we contribute the visual analytic framework VALUE which researchers can use to evaluate the utility of the joined datasets in a transparent manner (Section 5.5). Finally, we evaluate the algorithm and the VALUE framework through a usage scenario that helps demonstrate their efficacy through real-world datasets (Section 5.6).

5.2 Related Work

Evaluating the utility of joined open datasets has been a topic of considerable research for various use cases [140, 203, 26]; and, there is a growing need for developing robust metrics to quantify the usefulness of these joined datasets. Some research works discuss the quality of a dataset based on either the structure of the data or its content and then comment on improving its utility. For example, Ballou et al. first discuss the quality of data based on its completeness and/or consistency [204]. This paper proposes measuring completeness based on the presence of all elements and consistency as uniformity across comparable datasets, followed by a trade-off analysis

between these metrics to achieve the highest possible utility under a budget. But these metrics alone may not be sufficient to guide the selection of the most useful pair of joinable datasets from a large pool of open datasets.

Several other works have explored the challenge of balancing privacy and utility in datasets. For example, Kenneally and Claffy proposed the Privacy-Sensitive Sharing (PS2) framework to mitigate privacy risks while achieving utility goals when releasing datasets [205]. PS2 consists of components such as authorization, transparency, and access limitations that can help balance the privacy and utility aspects of released datasets. Bhumiratana and Bishop developed an ontology-based framework that enables formal and automatic communication between data collectors and users to ensure privacy-aware sharing of datasets, despite maintaining the utility of these datasets [206]. That said, privacy concerns may not always be relevant in evaluating the utility of joined datasets, especially when joining datasets about non-human objects. Moreover, while Noshad et al. proposed the Data Value Metric (DVM) to assess the information content of large datasets for augmentation in specific domains, this approach is limited to evaluating the utility of a single dataset rather than a joined open dataset [207].

Recent research has focused on different approaches for identifying joinable tables in large data lakes. For instance, Zhu et al. developed the JOSIE algorithm, which uses a set similarity search approach with a cost model to enhance performance over large data lakes [202]. However, an entirely automated approach may overlook the nuances of a human-centered approach, which is the focus of our work. Gong et al. developed the Niffler architecture, which finds joinable data tables over pathless table collections without join information [208]. But this approach does not enable the user to triage candidate datasets based on their utility. On the other hand, WarpGate, a semantic join discovery method implemented in Sigma workbooks, first indexes dataset columns and tries to find other datasets with similar columns [209]. Still, it

provides a score about joinability without a transparent explanation and options for exploration for the reasons behind it, which we attempt to explore through our visual analytic framework. Our work is comparable to the PEXESO framework by Dong et al., which converts the dataset columns into high-dimensional vectors and computes the similarity between these vectors to identify joinable tables [167]. Nevertheless, it does not quantify the utility of joining these datasets, which we attempt to do through the utility metric, which a researcher can transparently evaluate in order to update its components based on their background knowledge and expertise.

5.3 Understanding Join Scenarios

Understanding the various ways in which two datasets can be joined and the adaptability of a utility metric to different join scenarios is crucial for researchers seeking to gain insights from linking open datasets. Joining can be achieved through intersection, union, master join, or concatenation, each with different implications for the resulting dataset and its utility. The granularity of records, such as individual or aggregated levels, can also impact these join scenarios. In this section, we delve into these different scenarios and how they can influence the utility metric.

5.3.1 Intersection join

An Intersection join can be defined as the process of joining two datasets and keeping only those records that have matching values in both datasets for a specific combination of the join key attributes. This is one of the most common types of join encountered, also known as Inner join. Let's see an example of Intersection join.

Suppose we have two datasets, D1 (school records) and D2 (juvenile criminal activity records). A snapshot of D1 and D2 have been shown in Figure 5.1a and 5.1b respectively. Joining datasets D1 and D2 based on common attributes age, race, sex, and zip, we observe that there is only 1 common record of age 14, race Asian, gender F and zip 10012 (Figure 5.2a). We also observe extra information about this

age	race	sex	zip
10	Black	M	10012
14	Asian	F	10012
12	White	F	10011

a Dataset D1

age	race	sex	zip	crime
14	Asian	F	10012	larceny
17	Black	M	10013	theft
11	White	M	10021	battery

b Dataset D2

Figure 5.1 Snapshots of open datasets: (a) Dataset D1 shows the school records while (b) Dataset D2 shows the records of a juvenile criminal activities dataset.

individual that this individual has committed larceny. Thus, given the dataset D1, we can follow this process to identify other datasets that can be useful when joined with D1:

- Find datasets that have attributes common with that of D1 (like age, race, gender, and zip)
- Find if the records are similar. Since we need exact matches, we need to find a higher degree of similarity.
- Next, check if there is any other sensitive attribute revealed.

During the analysis of this join scenario, we discovered that record similarity and common attributes play crucial roles in determining the utility metric. It also became apparent that Intersection join is only practical for datasets with similar records, thereby enabling us to recommend pairs of datasets with high utility scores for Intersection join. This also highlights the need to set a defined range for the utility score to classify it as either "high" or "low".

a

Join Results

age	race	sex	zip	crime
14	Asian	F	10012	larceny

b

age	race	sex	zip	crime
10	Black	M	10012	NA
14	Asian	F	10012	larceny
12	White	F	10011	NA

c

age	race	sex	zip	crime
10	Black	M	10012	NA
14	Asian	F	10012	larceny
12	White	F	10011	NA
17	Black	M	10013	theft
11	White	M	10021	battery

d

age	race	sex	zip	crime
10	Black	M	10012	NA
14	Asian	F	10012	NA
12	White	F	10011	NA
14	Asian	F	10012	larceny
17	Black	M	10013	theft
11	White	M	10021	battery

Figure 5.2 Results from the Join Scenarios: (a) Intersection join (b) Master join (c) Union join and (d) Concatenation

5.3.2 Master join

Master join can be defined as the process of joining two datasets and keeping the records of either of the datasets and updating values or adding new attributes for those records which have matching values in both datasets, for a specific combination of the join key attributes. It is also known as Left or Right join in the SQL join parlance. This join is mainly useful when we intend to find extra information about the common records between two datasets.

If we perform a Master join on datasets D1 (Figure 5.1a) and D2 (Figure 5.1b), we would get an output similar to Figure 5.2b. Here, all the records from dataset D1 are retained, and the value for the new attribute (i.e., crime) has been updated.

Given dataset D1, the process of finding datasets for Master join is similar to that of Intersection join. Master join is preferred when the datasets have some similar records, and either dataset is selected as the primary one. Though the primacy has to be a user input, considering similarity as an essential constituent of the utility metric, we can say that Master join can be recommended when a pair of datasets have a medium range of utility score.

5.3.3 Union join

A Union join can be defined as the process of joining two datasets, keeping the records of both datasets and updating values for the common records, for a specific combination of the join key attributes. It is also known as Full join in the SQL join parlance. This join is mainly useful when we intend to keep the records from both datasets but update the values for the common records.

If we perform a Union join on D1 (Figure 5.1a) and D2 (Figure 5.1b), we will get an output similar to Figure 5.2c. Here, all the records from both datasets are retained, and the value for the *crime* attribute has been updated for the common record.

Given dataset D1, the process to find datasets for Union join is also similar to that of the other joins. However, unlike Master joins, Union join does not need a primary dataset since all the records will be retained. During our analysis, we realized that when there is a medium to low similarity between the records of datasets, it could be appropriate to consider a Union join. It is noted here that a Union join can only be performed when datasets have the same granularity. If the granularity is mixed, like having one individual and one aggregated record-level dataset, a Union join wouldn't make sense as it would create a joined dataset with mixed granularity.

5.3.4 Concatenation

Concatenation can be defined as the process of combining two datasets and keeping all records. Unlike Union joins, no attribute value is updated in this case.

If we perform a concatenation on D1 (Figure 5.1a) and D2 (Figure 5.1b), we will get an output similar to Figure 5.2d. Here, all the records from both datasets are retained as it is.

Given dataset D1, the process of finding datasets for Concatenation is also similar to that of other joins. But unlike Union join, Concatenation can still be

performed if there are some common attributes and no similar records. Thus, a low utility score can indicate a scenario for a Concatenation.

5.4 Calibrating Utility

Characterizing the join scenarios helped identify factors that need to be considered for calibrating the eventual utility of the join outcomes. In this section, we first summarize these factors and then we describe the algorithm.

5.4.1 Key factors impacting utility

Given a dataset D1, we observed that the following factors could be used to quantify the utility of joining it with another dataset:

Shared attributes in a dataset pair: The number of shared attributes between a pair of datasets is one of the important factors for determining the utility of the joined dataset. If two datasets do not share any shared attribute, there is no benefit in joining them through any join.

Degree of similarity between the records of the shared attributes: The degree of similarity can be an indicator of the utility of the joined datasets. We observed that datasets with similar records are useful while performing the joins, while datasets without any similar record can be used for concatenation.

Number of known shared attributes generally used for linking: Through our prior experience, we have observed that certain attributes are commonly employed to join datasets. We start with a list of known attributes like age, gender, race, and location. However, users can update this list based on their background knowledge and expertise. It also serves as a feedback mechanism in our human-in-the-loop approach, thus enabling the user to modify the inputs and transparently evaluate the utility of joining datasets.

While exploring other factors, we hypothesized that the number of exact matches between datasets would determine the join type. But after conducting some

experiments, we found that this hypothesis didn't always hold true. For example, even a single common record between datasets D1 and D2 could lead to a meaningful Intersection join, revealing sensitive information about an individual. Therefore, we decided not to incorporate it as a factor in our algorithm.

5.4.2 Utility metric

Algorithm 1 Utility Metric Algorithm

Require: Datasets D_1, D_2

Require: User supplied list of attributes generally used for linking (agl)

Require: $cutoffLength \leftarrow 200$

1: $f(D_i) \leftarrow$ attributes of D_i

2: $sa \leftarrow f(D_1) \cap f(D_2)$

3: $sa\_ratio \leftarrow |sa|/\{f(D_1) \cup f(D_2)\}$

4: $agl\_ratio \leftarrow (agl \cap sa)/|agl|$

5: $simNum, simCat \leftarrow [], []$

6: **for** each $attr$ in sa **do**

7: $Z_i \leftarrow dropNA(D_i.attr)$, where $i = 1, 2$ ▷ Keep only values

8: **if** $type(attr) =$ "numeric" **then**

9: $Z_i \leftarrow Z_i[: cutoffLength]$, where $i = 1, 2$

10: $sim \leftarrow cosineSimilarity(Z_1, Z_2)$

11: $AddItem(simNum, sim)$

12: **else**

13: $Z_i \leftarrow sort(Z_i, ascending)$, where $i = 1, 2$

14: $Z_i \leftarrow Z_i[: cutoffLength]$, where $i = 1, 2$

15: $C \leftarrow$ all Z_1-Z_2 combinations with one element from each

16: $temp \leftarrow []$

17: **for** each $comb$ in C **do**

18: $sim \leftarrow InDelSimilarity(comb[0], comb[1])$

19: $AddItem(temp, sim)$

20: **end for**

21: $simMean \leftarrow Mean(temp)$

22: $AddItem(simCat, simMean)$

23: **end if**

24: **end for**

25: $sim\_ratio \leftarrow Average(Mean(simNum), Mean(simCat))$

26: $w \leftarrow [20, 20, 60]$ ▷ Weights

27: $UtilityScore \leftarrow (w[0] * sa\_ratio) + (w[1] * agl\_ratio) + (w[2] * sim\_ratio)$

Algorithm 1 outlines the logic for our proposed utility metric. It is calculated as the weighted sum of three scores: *sa_ratio*, *agl_ratio*, and *sim_ratio*, reflecting the factors we identified as essential in calibrating the utility of joining datasets. Specifically, *sa_ratio* represents a normalized count of the shared attributes (sa) present while *agl_ratio* represents a normalized count of the attributes generally used for linking (agl) present in the shared attributes between datasets D_1 and D_2. To ensure consistency, each of these scores has been normalized to return a value between 0 and 1.

sim_ratio quantifies the similarity between the values of the shared attributes of the datasets. If all the values for a shared attribute are numeric, we calculate their cosine similarity using Python's scikit-learn package [210, 211]. However, if the values are categorical, we first generate all possible combinations of string values by selecting each value from records of the categorical attribute of each dataset. Then we calculate the similarity between each combination string using normalized InDel similarity from Python's Levenshtein package [212]. InDel distance is an edit distance between two

strings that calculates the number of insert/delete operations required to convert one string to another. The time complexity is $\mathcal{O}(m * n)$, where m and n are the number of characters in each string. This distance is then normalized over the maximum possible distance between two strings of size m and n, respectively. The normalized InDel similarity is then calculated as $1 -$ (normalized InDel distance). Finally, we compute the average of the categorical and numerical attributes' similarities to arrive at sim_ratio.

We use an edit distance-based similarity calculation method for finding the similarity between each record string. This method is preferable over token-based or sequence-based similarity calculations since the order of records does not affect our results significantly. We have considered several candidate algorithms for calculating the similarity between strings, including Levenshtein [213], InDel [214], Jaro-Winkler [215], and Hamming distance [216]. Hamming distance overlays one string over another and finds the number of places where the strings vary. While this method is effective for comparing strings of equal length, it is not well-suited for our purposes since the strings in our datasets can vary in length. Levenshtein distance calculates the number of operations (insert/delete/substitution) required to convert one string to another. Jaro-Winkler distance is similar to Levenshtein, but the substitution operation for close characters is given less weightage than that of far characters. InDel is a similar algorithm, but only insert and delete operations are allowed. We decided to use the InDel algorithm for string similarity calculation over Levenshtein and Jaro-Winkler algorithms. This choice was based on the fact that, in our current context, the substitution of characters may not be a reliable indicator of the level of similarity or difference between two strings of various types.

The final $UtilityScore$ is the weighted sum of these ratios, where more weight is given to the similarity between the attribute records. This score ranges between [0,100], thus making it easier to categorize high and low similarity, implementing

the insights gained while characterizing the join scenarios (Section 5.3). For computational efficiency, we have set a cutoff limit of 200 records for columns while calculating their similarity. Though this does not affect smaller datasets, for larger datasets, we can remove this constraint based on the availability of computational resources.

5.5 Framework for Transparent Evaluation of Utility

The algorithm for the utility metric can be best evaluated when paired with visual analytic interventions that a researcher can use to explore different open datasets and the utility of joining them. In this section, we first define the tasks for the VALUE framework and then discuss the visual analytic solution required to implement this framework on a web-based interface.

5.5.1 VALUE framework

The foremost challenge while assessing the utility of joining open datasets is to compare and triage different dataset pairs based on the utility metric. After that, researchers need to update the metric by considering their background knowledge, expertise, and analysis of the joined datasets. Centered around these steps, the tasks of the VALUE framework are as follows:

T1*: Inspecting utility scores*: The joinable groups of datasets can be further analyzed by ranking each pairwise dataset combination according to their utility score. This task relates to triaging dataset pairs based on their utility score. A dataset pair with a higher utility score will be more useful when joined based on some common attributes than one with a lower utility score. By identifying the most useful dataset pairs, researchers can focus their efforts on those with the highest potential for generating meaningful insights.

T2*: Incorporating user inputs to utility score*: The utility score of a joined dataset is influenced by the attributes commonly used to link two datasets, and this

Figure 5.3 Inspecting utility of joining real world open datasets through the VALUE interface: (a) A researcher selects a cluster of joinable open datasets based on relevant keywords. (b) Then all possible pairwise combinations of datasets are presented for the transparent inspection of the utility scores. Dataset pairs are ranked based on the utility score, and the user-selected attribute (*race*) present in the common attributes is highlighted for each pair. (c) Finally, the researcher can join the most useful pair and analyze the result through color-coded record categories. Numerical attributes are colored through an orange interpolation, while categorical attributes with less than ten categories are assigned distinct colors, and those with more than ten categories are colored through a grey interpolation.

evaluation can benefit from a human-in-the-loop approach. While we begin with a preliminary list of such attributes, a researcher can supplement this list based on their background knowledge and expertise. This task relates to the modification of the list of attributes generally used for linking, which can affect the utility score and ultimately lead to a meaningful join operation. By involving human expertise and feedback, we can ensure that the list of attributes generally used for linking is comprehensive and effective in capturing the most important aspects of the data.

***T3**: Analyzing joined records*: After joining the datasets, a researcher can perform a detailed analysis of the joined records to determine their utility. This task is necessary to extract valuable insights from the linked data and is essential for the success of a human-centered linked data analysis framework.

5.5.2 Visual analytic solution

The initial objective of the VALUE framework is to identify groups of joinable open datasets, and it is accomplished using two key visual analytic components. The first component is a search box that allows the user to filter datasets based on relevant keywords. The second component is a high-dimensional projection of the datasets based on their similarity in their attribute space (Figure 5.3a). In order to achieve this, we first transform the dataset attributes into high-dimensional word embedding vectors. These vectors are then projected onto their two-dimensional (2-D) space using the t-SNE dimensionality reduction algorithm [183]. Then we apply the DBSCAN algorithm to identify and group datasets with similar attributes into clusters [186]. Datasets that belong to the same cluster are color-coded for easy identification. Furthermore, each cluster is ranked based on its intra-cluster distance using the Silhouette coefficient [191], and individual datasets within the cluster are labeled accordingly. This approach allows the researchers to comprehend the relationships between datasets and identify joinable groups.

Once a researcher selects a group of joinable datasets, all possible pairwise combinations of datasets are displayed for further inspection (Figure 5.3b). Each dataset pair is visually represented using a combination of items, such as the dataset names, rectangular boxes showing the common attributes between these datasets, and their utility score. The utility score is represented using a horizontal green bar where the color green represents the score in a range of 0-100. This abstraction provides a convenient way for the researchers to understand the scores at a glance, but they can also obtain exact score information by hovering over the bar. The choice of the color green is purely for semantic reasons. The similarity between the records of common categorical and numerical attributes is also shown with two vertical green bars. These bars' orientations have been reversed to differentiate them from the main utility score. Thus, this design aids the transparent evaluation of the utility scores **(T1)**. Furthermore, this view also enables the researcher to augment the list of the attributes generally used for joining. If any of these attributes are present in the common attributes, they are highlighted in a distinct color (royal heath) to indicate their significance. This human-in-the-loop approach helps to improve the utility score based on the inputs from the researcher **(T2)**.

As learned during the characterization of join scenarios, we recommend Intersection join for dataset pairs with high utility scores. To facilitate this, the button for Intersection join is highlighted, but the researcher has the option to choose any other type of join. The joined datasets are visualized through a customized Navio implementation, where each attribute is represented by a stacked bar chart displaying the distribution of different categories for that attribute (Figure 5.3c) [217]. For a numerical attribute, the records are represented using a sequential scheme of colors. The null values, shown in light pink, help to understand the completeness of the results. This colored categorization of the joined dataset's records helps a researcher understand its composition and analyze them for utility **(T3)**. Users can

also download the joined dataset for further investigation. The web-based interface has been developed using a combination of Python and Flask for the backend and Node.js, React.js, and JavaScript for the frontend.

5.6 Usage Scenario

The performance of the algorithm for utility metric and the VALUE framework can be evaluated in multiple ways. A systematic review by Isenberg et al. observed that Qualitative Result Inspection is one of the most popular evaluation methods for algorithms and visualization interfaces [218]. Hence, in this section, we describe a usage scenario to demonstrate the how the visual analytic interface that embeds the utility metrics can help in distinguishing between the highly usable and the least usable pairwise join outcomes.

Consider a scenario where a researcher at a government laboratory is analyzing local election results obtained from open datasets to gain insights that could inform policy decisions or contribute to a broader understanding of the political landscape in the area. The findings of the study could be crucial for stakeholders such as policymakers, government agencies, or local communities in formulating informed decisions. She began by browsing several county-level open data portals to obtain the necessary data. However, she found it challenging to determine which datasets to combine to form a complete picture of the election results. In search of a solution, she turned to the VALUE interface. After conducting a search on elections, the interface generated several clusters of data related to election results. To make her selection, she carefully analyzed the projection plot and ultimately chose the first cluster (Figure 5.3a).

This action generated all the possible pairwise dataset combinations from this cluster and ranked them according to their utility score (Figure 5.3b). The researcher analyzed the dataset pairs and observed that the datasets *2010 General - Election*

Results by precinct (complete eCanvass dataset) and *2010 Primary - Election Results by precinct (complete eCanvass dataset)* have a high utility score of 82.77 **(T1)**. These are the datasets for the general and primary election results of 2010 from King County, WA. Since this pair has a high utility score, the interface suggested an Intersection join between the datasets. She also observed that this pair included one attribute (race) that was included in the default list of generally used attributes **(T2)**. Though she did not update this list, she selected all the attributes and performed an intersection join.

On joining these datasets based on all the common attributes, the researcher observed that the joined dataset contains 162, 977 records (Figure 5.3c). She analyzed these records using the VALUE interface and understood that the joined dataset gives her the combined election results for all the candidates at each precinct, both at the primary and general election levels **(T3)**. She further downloaded the joined dataset and saved it for her research purposes.

Further, the researcher was curious to understand if the utility metric could distinguish between the most useful and the least useful dataset pairs. Hence, she selected the lowest ranked dataset pair: *2010 Primary - Election Results by precinct (complete eCanvass dataset)* and *Election night precinct results - November 2018*. Joining them based on the common attributes ['race', 'precinct'] yielded no record. Thus, the researcher concluded that the utility metric, when used in conjunction with the VALUE framework, can help to find joinable and useful datasets from the open data ecosystem.

5.7 Discussion

The utility metric can be considered a novel method that can be used to assess the utility of joining open datasets with a human-centric perspective. In our continuous efforts to improve and refine the algorithm behind the utility metric, we aim to unlock

even greater insights into the potential of joining open data. We also plan to use the outcomes from the utility metric to train a machine-learning model to classify the usefulness of the joins.

The insights gained from our analysis of the join scenarios represent a crucial foundation for this work. By leveraging the interface for the VALUE framework, we could put some of these lessons into practice, emphasizing the critical role of visual analytic interventions in solving this problem. Although the current interface prototype is designed to work with approximately 400 open datasets, our internal testing has indicated that it can be scaled up significantly. Also, while we did need to implement a cutoff length for larger datasets, we are currently exploring strategies to overcome this limitation, such as increasing our computational resources. Additionally, we are also working on a workflow that can regularly fetch datasets from different sources and integrate them with the VALUE framework, thus enabling us to keep pace with the ever-evolving landscape of open data.

We recognize that there is always room for improvement in the interface components of the VALUE framework, and we are committed to incorporating feedback from a diverse range of users. To this end, we plan to conduct case studies with domain experts and undertake more controlled user studies that will enable us to collect valuable feedback about the interface and the algorithm.

5.8 Conclusion

The utility metric algorithm, presented in this chapter, is a first step towards quantifying the utility of joining open datasets. It considers multiple factors like the similarity between records, shared attributes, and a user-supplied list of attributes to develop a score that can help identify the most useful pair of datasets from a group of joinable datasets. The lessons learned during this development also helped develop the VALUE framework, which, when used in conjunction with the web-based interface,

helps in the transparent evaluation of the utility score. This human-in-the-loop approach helps researchers, data scientists, and analysts to make more informed decisions and leverage the full potential of open datasets.

CHAPTER 6

LINKLENS: WORKFLOW FOR BALANCING PRIVACY AND UTILITY FACTORS IN MULTI-WAY JOINS

6.1 Introduction

Open datasets across diverse domains such as health, economics, and politics are readily accessible in open data repositories. Users frequently aim to join datasets from various domains to extract insights spanning these diverse areas. For example, aggregating election results from different regions can reveal voting trends. The joining process can involve various datasets and different types of joins; for instance, election result datasets from multiple years may initially be concatenated and then intersected with infrastructure project datasets to analyze development trends under different elected officials. However, the sheer volume of potential combinations can overwhelm users. For instance, with a group of 150 datasets, there could be over 11,000 pairwise combinations and over 550,000 three-way combinations (Figure 6.1). Moreover, each three-way combination can be arranged in three ways, leading to different outcomes, which effectively creates a combinatorial explosion.

Moreover, such dataset join can risk identity or attribute disclosure, where a data subject's exact identity or identifiable attributes may be revealed. For instance, despite the Australian Department of Health releasing de-identified data about 2.9 million patients, researchers managed to re-identify patients and their doctors within a few months by leveraging other open demographic information [13]. In another instance, 91% of all taxis operating in NYC were identified using de-identified NYC taxi open data and other open taxi datasets [140]. Similarly, in our previous work, we showed several instances where individuals were identified and sensitive information was revealed by linking criminal record datasets [25]. Additionally, practitioners commonly adhere to the "release-and-forget" model, where open datasets, once

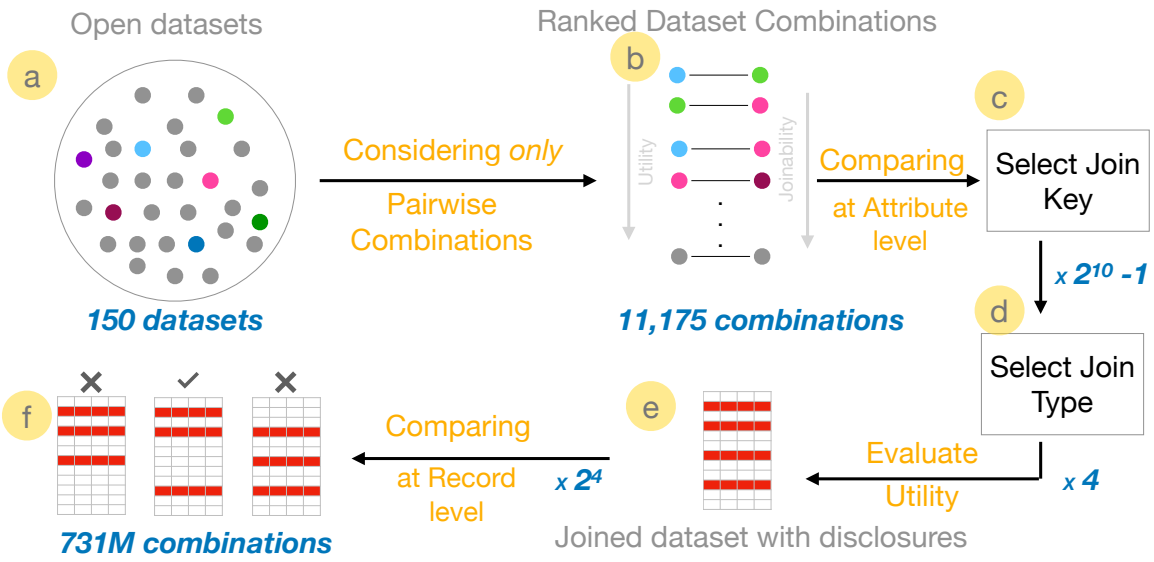


Figure 6.1 Process Overview: (a) - (f) illustrates the process of discovering joinable open datasets and evaluating the balance between utility and privacy. Even with basic assumptions and considering only pairwise combinations, a mere 150 datasets can yield over 731 million combinations. Including multi-way join can well lead to a combinatorial explosion. Visual analytics can aid in navigating this complexity, making it easier to balance privacy and utility factors when joining open datasets.

released, are not systematically reviewed for potential disclosures in light of newly released datasets [51]. Consequently, consistent scrutiny of the privacy implications for data subjects when combining open datasets is crucial.

Balancing utility and privacy encompasses a broad spectrum of scenarios, from scenarios where dataset utility is maximized without regard to disclosures to those where utility is minimized by removing all potential disclosure records. Users can opt for either extremes or any intermediate scenario. This aligns with the broader concept of trade-off scenarios, defined as "How much achievement on objective 1 is the decision-maker willing to give up in order to improve achievement on objective 2 by some fixed amount?" [219]. As evident from the earlier examples, employing different join types (e.g., intersection, union, or master join) compounds this combinatorial complexity, thus making it difficult for a user to compare all the scenarios and make

an informed decision. As depicted in Figure 6.1, with basic assumptions, just 150 datasets can generate approximately 731 million combinations.

Visual analytics can play an important role in addressing the challenges posed by combinatorial complexity in dataset exploration and facilitating the analysis of disclosures. For instance, UrbanForest utilized a heatmap-based design to illustrate the relationships between different datasets and attributes, aiding users in selecting relevant datasets more efficiently [41]. PRIVEE developed a visual analytic workflow to identify joinable groups of open datasets, triage them based on their joinability risk, and finally evaluate disclosure risk at the record-level [26].

In this context, we first contribute a novel human-in-the-loop visual analytic system, LinkLens, to systematically balance the privacy risks and the utility of joinable open datasets. Users, like data owners and custodians, can navigate joinable datasets from over 100 open data portals, analyzing their utility while simultaneously inspecting them for potential disclosures. This inspection facilitiated through our second contribution, the development of utility and joinability risk scores that is specifically designed for multi-way joins. Visual analytic interventions in LinkLens are designed to guide users in making context-aware decisions about risk tolerance and the perceived utility of joinable datasets, ensuring the discovery of useful open datasets with reduced risks of disclosures. In this chapter, we begin by a discussion of the goals and the visual analytic tasks for LinkLens. Next, we describe how we map these goals with the interface design and demonstrate the efficacy of LinkLens through a usage scenario.

6.2 Visual Analytic Goals and Tasks

Manually discovering, combining, and evaluating datasets for both utility and disclosure at the metadata and record levels can be a challenging task. In contrast, during our research, we realized that completely automating this process may be

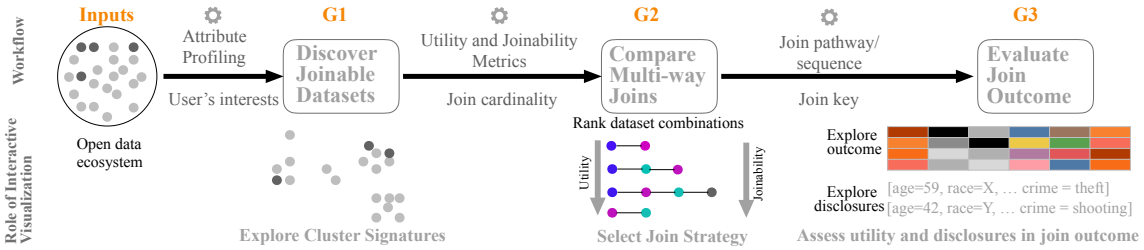


Figure 6.2 LinkLens workflow: This workflow enables users to discover joinable open datasets aligned with their interests, compare multi-way join options, and assess the outcomes based on utility and potential disclosures. Interactive visualization enhances this process by guiding users through each step of the workflow.

practically infeasible as human intervention is required at multiple stages of dataset exploration, interpretation, and evaluation. To address this challenge, we developed LinkLens, a visual analytic workflow designed to help the users balance privacy and utility while performing multi-way joins of open datasets (Figure 6.2). In this section, we describe the LinkLens workflow through the high-level goals and visual analytic tasks required to achieve these goals.

G1: Discover joinable datasets: The first goal of the LinkLens workflow is to identify open datasets that can be combined to extract valuable insights. This involves considering various join combinations, including pairwise, three-way, four-way, and more complex scenarios. Additionally, some datasets may exhibit transitive joinability, meaning they lack direct shared attributes but can be connected through a third dataset. These situations increase the complexity of this goal, which may be achieved through these tasks:

T1: *Explore dataset cluster signatures:* Since the presence or absence of shared attributes significantly influence dataset joinability, clustering datasets by their attribute space is a promising approach for finding joinable open datasets. Datasets with similar attribute spaces are more likely to be joinable. However, understanding the specific nature of attributes within each cluster is essential for identifying potential join candidates. This task involves analyzing the distribution of attributes within

each cluster, enabling users to identify datasets that closely align with their research objectives.

T2: *Select attributes of interest:* In our previous research, we have identified a list of attributes which are frequently used to find datasets that can potentially reveal sensitive information. These quasi-identifiers, such as age, race, gender, location, while individually insufficient for unique identification, can collectively help identify a data subject, thus leading to identity and attribute disclosure. If selected by users, these attributes can highlight datasets at higher risk of disclosure when joined with others. By considering such dataset combinations, data defenders can evaluate the trade-offs between the utility and privacy of a dataset join.

T3: *Select joinable dataset combinations:* Understanding the joinability between datasets within and across clusters can help in selecting dataset combinations that effectively support user objectives. This task revolves around combining two or more datasets, which can broaden the scope of the final joined dataset, potentially providing a larger volume of data. The number of datasets selected for joining is a critical factor in generating meaningful and informative results, encompassing data from multiple sources and potentially diverse domains.

G2: Compare multi-way join options: Joining two datasets results in a single combination, as the order of the join does not affect the outcome. But joining three datasets can yield three distinct permutations since order of the join can lead to different outcomes. The number of possible permutations increases with the number of datasets involved. Thus, for higher-order multi-way joins, it is important to compare different join options which can be achieved through these tasks:

T4: *Explore join cardinality:* While adding more datasets to a join can potentially increase the volume of data, it also carries the risk of diluting data quality due to the inclusion of irrelevant information. Therefore, it is important to carefully evaluate different join options before combining the datasets. This task involves

comparing the benefits and drawbacks of 2-way, 3-way, and n-way joins to determine the most appropriate approach for the user's specific needs.

T5: *Triage multi-way combinations based on utility and joinability risk:* Consider a four-way join involving datasets from a specific domain with a few shared attributes. Even within this scenario, different join combinations may yield varying levels of utility and joinability risk. Therefore, it is essential to carefully evaluate and prioritize various multi-way join options before selecting a join strategy.

T6: *Select join strategy:* A group of datasets can be joined in various ways depending upon the type of join. For example, in a four-way join, first two datasets can be joined using intersection join on attributes a1 and a2. The resulting dataset can then be joined with a third dataset using a union join on attributes a3 and a4, based on user requirements. This process can continue with the fourth dataset, employing different join types and shared attributes. Modifying any join type and shared attribute may result in a different join outcome. Hence, selecting a proper join strategy becomes a very important task in this workflow.

G3: Evaluate join outcome: The ultimate goal of the LinkLens workflow is to evaluate the join outcome and assess its suitability for further use. Analyzing the records to determine their utility in fulfilling user requirements is an important takeaway from this workflow. Moreover, it is equally important to evaluate these datasets for potential privacy risks before exporting them. This goal relates to the balancing act between utility and privacy of the joined dataset and can be achieved using the following tasks:

T7: *Assess utility of join outcome dataset:* Users seek to determine if a specific join strategy results in a dataset useful for their research needs. This can be done by analyzing the distribution of the records in this dataset and ascertaining if they really generate some useful information. This task entails evaluating the utility of the join outcome dataset based on the data distribution at the record level.

T8: *Inspect join outcome for possible disclosures:* Any evaluation of the join outcome would be insufficient without examining it for potential disclosures. Exporting a dataset with potential vulnerabilities can lead to sensitive information leaks within the user's pipeline. To mitigate these risks, this task focuses on examining the join outcome for potential disclosures at the record level.

6.3 Design Methodology

The design of LinkLens is motivated by the transparent explanation and the evaluation of the utility and risk assessment process. In order to do that, we implement a web-based interactive interface that allows users to discover joinable open datasets, compare different join options and evaluate the joined datasets for utility and possible disclosures. This interface is developed using Python and Flask for the backend services while the front-end was developed using Javascript frameworks like React.js and D3.js. In this section, we provide an overview of the design requirements required for realizing the visual analytic goals and tasks of the LinkLens workflow.

Datasets profiling at the metadata level: As a part of our previous research, we collected approximately 5400 open datasets containing various combinations of popular quasi-identifiers like age, gender, race, location and others. After multiple levels of filtering, we selected a set of 426 datasets which contain data about human subjects and could be vulnerable to sensitive information disclosure. These datasets serve as the input to LinkLens and we have profiled them based on their record granularity, number of records, and attributes. The landing page for LinkLens allows users to select attributes of interest from the quasi-identifiers and filter results by record granularity.

Exploring joinable datasets: To fulfill G1, LinkLens clusters joinable datasets and represents them using a set of visualizations including a projection plot and a group of bar charts. Datasets within the same cluster may have stronger

joinability, making helpful for identifying potential join candidates. However, mouse interactions enable users to explore joinability both within and across clusters, ensuring the search is not limited to a single cluster. Although LinkLens automates dataset grouping, the visualizations, especially the bar charts for each cluster, help users transparently understand the reasons influencing cluster formation.

Triaging join options: LinkLens automatically calculates the possible join permutations based on the number of datasets selected. These permutations are ordered by utility, allowing users to avoid manually reviewing all options before choosing a join strategy. Visual cues indicate the utility and joinability risk for each permutation, as well as the shared attributes between datasets. This automated, human-in-the-loop design helps users to evaluate a large number of join options and make an informed decision about their join strategy.

Inspecting joined dataset: After selecting a join strategy using a specific permutation of datasets, join key and join type, users need to inspect the joined datasets at the record level to ensure they meet their research needs. LinkLens shows the distribution of categories for each attribute present in the joined datasets in a compact format, allowing users to quickly understand the data distribution. A separate table shows the possible disclosures so that the users can review them before exporting the data for further use. These visual analytic cues help users to compare the joined dataset's utility against the possible disclosure risk and make an informed decision.

6.4 Discover Joinable Datasets (G1)

Users need to understand the degree of joinability between datasets in order to make a decision about selecting datasets for join. Hence, the design requirements for addressing tasks T1, T2 and T3 are to develop clustering methods and use visualization cues that are responsive to user-defined attributes of interest along with

transparency in explaining cluster signatures. This enables the users to create mental model of the joinability between different datasets and make a choice toward selecting groups of datasets of their choice. In this section, we first describe the clustering methods for finding joinable datasets and then describe how our design choices help users to understand the reason behind formation of these clusters.

6.4.1 Clustering methods for finding joinable datasets

Converting Data Attributes to Word Embeddings: As pointed out earlier, the joinability of two datasets depends on the shared attributes. Therefore, datasets with similar attributes are likely more joinable. But attribute names in open datasets are often noisy and inconsistent, making it difficult to perform a binary search for specific attributes. We address this by focusing on the idea that similar attribute names can reflect the semantic similarity among datasets sharing a similar context. To achieve this, we utilize a word-embedding approach that captures both joinability and semantic similarity. Word embeddings are real-valued, fixed-length, dense representations that capture lexical semantics [178, 179]. We transformed the data attributes into their word embedding form using Python's spaCy library, generating a vector representation for each dataset's attribute space [180]. Vectors with smaller distances between them indicate datasets with similar attributes and, thus, greater joinability. We use cosine similarity to measure the similarity between these vectors [181, 182].

Projecting the datasets and finding clusters: Each dataset is now represented by a vector with over 300 dimensions. However, comparing datasets in a 2-D or 3-D plot is difficult due to the high dimensionality. Therefore, we used the t-SNE dimensionality reduction algorithm to transform these high-dimensional vectors into two-dimensional representations [183]. A 2-D projection alone may not effectively reveal dataset groupings, so we tested clustering algorithms like KMeans [184],

DBSCAN [185, 186], Birch [187], and OPTICS [188, 189]. After a thorough analysis of cluster quality and density scores, we chose the DBSCAN algorithm. This method provided the most distinct and meaningful clusters in our experiments, enhancing our ability to identify joinable datasets.

Evaluating the clusters: There can be multiple groups of similar or joinable datasets, leading to the creation of several clusters. Assessing all these clusters can be difficult for a user, so we used cluster evaluation techniques to prioritize them. One such metric is the *Calinski-Harabasz Index*, which is defined as the ratio of between-cluster dispersion to within-cluster dispersion. Here, dispersion refers to the sum of squared distances between samples and their cluster's barycenter [190]. A higher score indicates better cluster separation and formation. We conducted an experiment to compare this metric with other metrics like the Silhouette Score [191] and the Davies-Bouldin Index [192]. The Calinski-Harabasz Index was selected because it efficiently guided users in identifying meaningful, joinable datasets.

6.4.2 Dataset joinability view

We designed Dataset Joinability View (Figure 6.3) to give users an overview of all available datasets, allowing them to explore signatures and patterns (T1) and select a set of joinable datasets (T3) that align with their interests. LinkLens automatically highlights the popular quasi-identifiers selected by the users (T2) in order to augment the decision making process. The components of this view are as follows:

Joinable groups of datasets: LinkLens represents the datasets using a 2-D projection plot where each dataset is represented by a grey circle. Datasets positioned closer to each other in this plot are similar in their attribute space. While the initial version of this plot used color to differentiate between the clusters of datasets, we now use only grey to denote datasets since they are already distinctively clustered (Figure 6.3a). Hovering over a dataset provides additional information, such

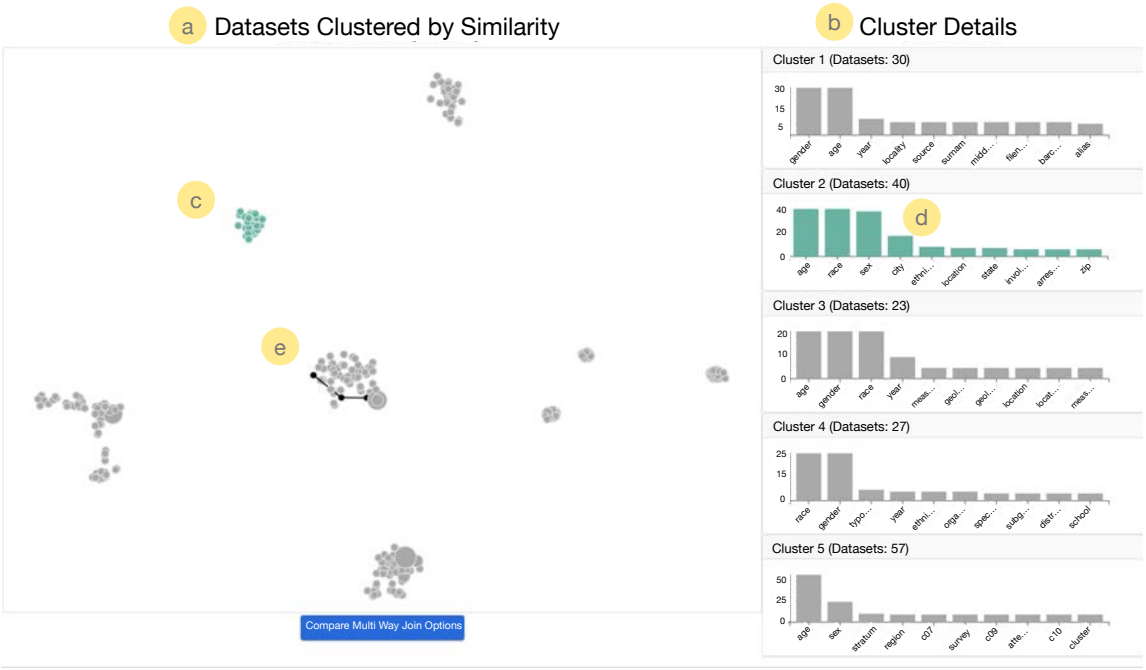


Figure 6.3 Dataset Joinability View: (a) LinkLens clusters available datasets based on their similarity in attribute space, and (b) bar charts display the frequency of common attributes, explaining the cluster formation. (c, d) Hovering over the cluster bar chart highlights the corresponding group of datasets (shown in dark green), and vice-versa. (e) Clicking on a dataset allows users to explore and select potential join pathways.

as the dataset name, the open data portal it belongs to, and popular domain tags from their open data portal. The size of the dots indicates the number of records in each dataset, helping users determine if the dataset is large enough for their research purposes. This information collectively helps the user select a starting dataset.

When a user hovers over another dataset, a dotted line appears between the two datasets, showing the number of shared attributes. This visual cue assists in selecting the next dataset for a multi-way join. Users can repeat this process multiple times to create a joinable dataset combination **(T3)** (black connected lines as shown in Figure 6.3e). That said, if the user selects specific attributes of interest, datasets containing these attributes are highlighted using a darker shade of grey, aiding the decision-making process based on their interests **(T1)**.

Transparent explanation of joinability: As mentioned earlier, the Calinski-Harabasz Index is used to rank the clusters, thus aiding in a more effective explanation

of the reasons behind their formation. Since the highest-ranked cluster in this method is the most closely-knit, the datasets within it are more likely to be joinable. Therefore, we prioritize this cluster for users, allowing them to explore its joinability first. Hence, for each cluster, we display a bar chart showing the most frequent attributes of that cluster (Figure 6.3b). This helps users understand the signature of a particular cluster and understand if aligns with their interest **(T2)**. Attributes of interest selected earlier are highlighted in the bar chart using the same darker shade of grey **(T1)**.

The projection plot and bar charts are interlinked through mouse interactions. Hovering over a bar chart temporarily highlights the datasets in that cluster using a distinct green color. Similarly, drawing a lasso around a group of datasets in the projection plot will highlight and focus on the relevant bar chart(s). For example, in Figure 6.3, Cluster 2 has been highlighted through the projection plot (Figure 6.3c) and the relevant bar chart has been highlighted on the right hand side (Figure 6.3d). All these interactions together contribute towards a better explanability of the reason behind the formation of the clusters.

6.5 Compare Multi-way Join Options (G2)

After selecting a set of datasets, the next step in the LinkLens workflow is to compare different join cardinalities (T4), rank the dataset combinations based on different metrics (T5) and then develop a join strategy (T6). This is done through the Join Comparison View where users can compare different dataset combinations of varying cardinalities and then select a join strategy that gives the highest utility with the lowest possible joinability risk. In this section, we first discuss the algorithms for generating dataset combinations of different join cardinalities, utility scores and joinability scores, followed by how all of them are used together with different visual analytic interventions to form the Join Comparison View.

6.5.1 Metrics for utility and risk comparison

Calculating the possible join combinations: A multi-way join is an operation that combines two or more datasets into a single, unified outcome. The order of combining datasets within a multi-way join can vary, thus, leading to multiple possible pathways to achieve the final outcome. For example, while joining datasets D_1, D_2 and D_3, the pathway ((D_1 X D_2) X D_3) and (D_1 X (D_2 X D_3)) would yield different outcomes, since the datasets in the parenthesis are joined first. To determine the number of distinct pathways for performing a multi-way join, we can frame the question as follows: "*In how many ways can you sequentially join N datasets, where the order within the innermost parentheses doesn't matter, but the order of subsequent joins does matter, and the order of joining two results of equal size doesn't matter?*"

We conducted experiments to determine the number of possible pathways for multi-way joins and found that it closely resembles the Catalan numbers [220, 221]. In this context, the number of ways to perform pairwise joins on N datasets where order matters is represented by the Catalan numbers. Mathematically, it is defined as:

$$Ct(N) = \frac{(2N)!}{(N+1)! * N!} \tag{6.1}$$

Extending upon this, we propose that the number of possible join pathways given a certain number of datasets (N; $N >= 3$) is as follows:

$$MW(N) = \frac{{}^{N}C_2 * Ct(N-1)}{2} \tag{6.2}$$

The logic behind this formula can be explained as follows: We start by selecting two datasets out of N to join, which can be done in ${}^{N}C_2$ ways. This initial step is crucial as every pathway starts with joining two datasets. After this initial join, we are left with $N-1$ units to join ($N-2$ original datasets plus the one resulting from the initial join). The Catalan number $Ct(N-1)$ counts the number of ways to parenthesize and

order these remaining $N-1$ joins, as each set of parentheses represents a pairwise join operation. However, the product of ${}^{N}C_2$ and $Ct(N-1)$ initially overcounts the distinct combinations because the order in which the two initial pairs are formed does not matter (for example, (D_1 X D_2) X (D_3 X D_4) is the same as (D_3 X D_4) X (D_1 X D_2)). To correct this overcounting, we divide by 2!, which accounts for the number of ways to arrange the two resulting datasets after the initial pairwise joins. A list of possible join pathways at different join cardinalities ($N = 2, N = 3, N = 4$ ...) can be found in the supplementary materials.

Calculating utility score and joinability risk:

Algorithm 2 Utility and Risk Metric Algorithm

Require: *pathways*: All possible pathways

▷ Example: [[(D_1 X D_2) X D_3], [D_1 X (D_2 X D_3)] ...]

▷ D_i is each dataset

Require: *aoi*: User supplied attributes of interest

1: $utilityScores, riskScores \leftarrow [], []$

2: **for** *index*, *pathway* in enumerate(*pathways*) **do**

▷ : Pathway : (D_1 X D_2) X D_3

3: $columnsNames, columnsData,$

4: $columnsTypes \leftarrow processPathway(pathway)$

5: $f(D_i) \leftarrow$ attributes of dataset D_i

6: $sa \leftarrow \bigcap_{i=1}^{n} f(D_i)$ ▷ $n = 3$ in this example

7: $allA \leftarrow \bigcup_{i=1}^{n} f(D_i)$

8: $saRatio \leftarrow |sa|/|allA|$

9: $aoiRatio \leftarrow |aoi \bigcap sa|/|aoi|$

10: $simRatio \leftarrow 0$

11: $riskScore \leftarrow 0$

12: $counter \leftarrow [0, 1, ..., n-1]$

13: **for** cr in counter **do**

14: **if** $cr \neq 0$ **then**

15: $simRatioCr, riskScoreCr$

16: $\leftarrow$ **scoreCalculationCr**$(0, ct, columnsNames,$

17: $columnsData, columnsTypes,$

18: $aoi)$

19: $simRatioCrMean \leftarrow \frac{\sum simRatioCr}{len(simRatioCr)}$

20: $simRatio \leftarrow simRatio + \frac{simRatioCrMean}{cr}$

21: $riskScore \leftarrow riskScore + \frac{riskScoreCr}{cr}$

22: **end if**

23: **end for**

24: $w \leftarrow [20, 20, 60]$ ▷ Weights

25: $utilityScore \leftarrow (w[0] * saRatio) + (w[1] * aoiRatio) + (w[2] * simRatio)$

26: $utilityScores \leftarrow utilityScores + [utilityScore]$

27: $riskScores \leftarrow riskScores + [riskScore]$

28: **end for**

return $utilityScores, joinabilityRiskScores$

In our previous work, we have developed algorithms to calculate the utility score [27] and joinability risk [26] for a pairwise combination of datasets. But in multi-way join, each join pathway may contain more than two datasets and the order of these datasets will determine both the utility and the joinability risk of the pathway. Therefore, we have extended and combined these algorithms to calculate the utility score and joinability risk for each possible pathway. The logic for this extended algorithm is outlined in Algorithm 2.

In this algorithm, we first process each pathway and extract the column names, column types and the column data for each attribute present in the datasets of the

Algorithm 3 Sub-function: scoreCalculationCr

Require: i, j: Starting and ending indices

Require: $columnsNames$: List of all column names

Require: $columnsData$: Dictionary containing data for each column

Require: $columnsTypes$: Dictionary containing data types for each column in each column

Require: aoi: User supplied attributes of interest

$subSet1 \leftarrow \bigcup_{k=i}^{j-1} columnsNames[k]$

$sharedAttrs \leftarrow subSet1 \cap columnsNames[j]$

$rw \leftarrow [50, 1]$ ▷ Risk weights

$aoiR \leftarrow (aoi \cap sharedAttrs)$ ▷ aoi Ratio

$riskScore \leftarrow (rw[0] * |aoiR|) + (rw[1] * (|sharedAttrs| - |aoiR|))$

A, B, $dataType \leftarrow [], [], []$

$saScores \leftarrow []$

for sa in $sharedAttrs$ **do**

    $A \leftarrow \bigcup_{k=i}^{j-1} columnsData[k]$

    $B \leftarrow columnsData[j]$

    if all($type(sa)$ = "numeric") **then**

        $sim \leftarrow cosineSimilarity(A, B)$

        $saScores \leftarrow saScores + [sim]$

14: **else**

15: $C \leftarrow \prod_{i \in A, j \in B}\{i, j\}$ ▷ Cartesian of A and B

16: $temp \leftarrow []$

17: **for** each $comb$ in C **do**

18: $sim \leftarrow InDelSimilarity(comb[0], comb[1])$

19: $temp \leftarrow temp + [sim]$

20: **end for**

21: $simMean \leftarrow Mean(temp)$

22: $saScores \leftarrow saScores + [simMean]$

23: **end if**

24: **end for**

return $saScores$, $riskScore$

pathways. These data are arranged according to the order of joins specified in the pathway. This is achieved using a tree-like structure, where the innermost datasets are the leaf nodes, and the subsequent datasets are higher-order nodes. More details about this tree processing mechanism are available in the supplementary materials.For each pathway, we calculate the shared attributes among the datasets and derive two ratios: the *saRatio* and the *aoiRatio*. The *saRatio* represents the percentage of attributes shared by all datasets compared to the total attributes available, while the *aoiRatio* indicates the percentage of attributes of interest present in the shared attributes relative to all possible quasi-identifiers. The third ratio, *simRatio*, is calculated through the Algorithm 3, quantifying the similarity between the values of the shared attributes. For each pathway, the *simRatio* at each level is a weighted sum, giving more preference to the innermost join. The final *utilityScore* is a weighted sum of these ratios, with more weight given to the similarity between attribute records.

This score ranges between [0,100], thus making it easier to categorize high and low utility.

In Algorithm 3, we first calculate the *riskScore* as the weighted sum of the number of quasi-identifiers and non-quasi-identifiers present in the shared attributes. The weights are calculated through an experiment involving different weight combinations and further detail can be found in the supplementary materials. To calculate the *simRatio* for the *utilityScore*, we first divide the data into two parts: A and B. A consists of the union of all column values for a specific shared attribute across all datasets except the last one, while B contains the same for the last dataset. For example, in the pathway (D_1 X D_2) X D_3, the first iteration calculates the *simRatio* between each shared attribute(sa) of D_1 and D_2, where $A = D_1[sa]$ and $B = D_2[sa]$. If the attribute type is numeric, we calculate the cosine similarity between A and B using the Python scikit-learn package [210, 211]. On the other hand, if the attribute type is categorical, we first generate all possible combinations of string values by selecting each value from A and B. Then we calculate the similarity between each combination string using normalized InDel similarity from Python's Levenshtein package [212]. InDel distance is an edit distance between two strings that calculates the number of insert/delete operations required to convert one string to another and the time complexity is $\mathcal{O}(m * n)$, where m and n are the number of characters in each string. This distance is then normalized over the maximum possible distance between two strings of size m and n, respectively. The normalized InDel similarity is then calculated as $1 -$ (normalized InDel distance). As per Algorithm 3, these values are stored in an array and returned along with the *riskScore*. In Algorithm 2, we first calculate a mean of these values as $simRatioCtMean$ and a weighted version of it is added to the *simRatio* variable. In this case, since this is the first iteration for this pathway, the $simRatioCtMean$ is divided by 1 (basically the same value) and added to *simRatio*. A similar approach is applied to the *riskScore*. In the

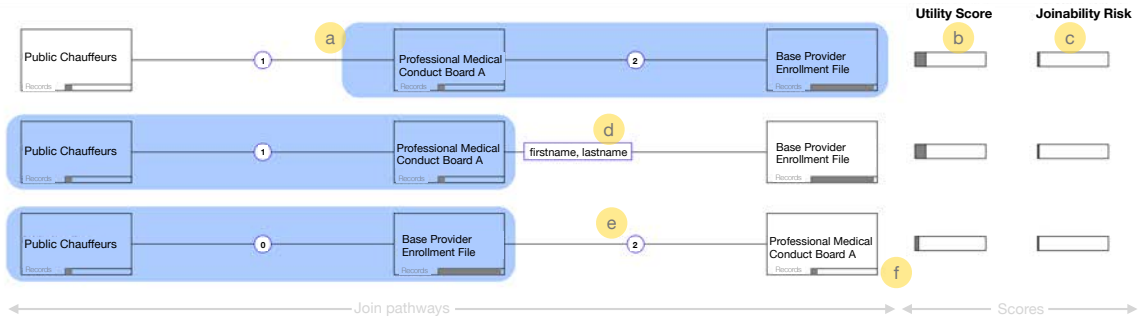


Figure 6.4 Join Comparison View: LinkLens allows users to compare different join pathways, with (a) the join order in a selected pathway highlighted in blue and (b, c) the utility and joinability risk for each pathway represented by grey bars. (d) Shared attributes between datasets are shown as boxes on the connecting lines, while (e) the total number of shared attributes is displayed using a circle and text view to provide a high-level overview. (f) Small grey bars within each dataset indicate the record count relative to others in the pathway, helping users assess whether they are worth joining.

second iteration, we now consider $A = D_1[sa] \cup D_2[sa]$ while $B = D_3[sa]$. Similar process is undertaken to calculate the $simRatioCtMean$ through the Algorithm 3. However, this time, it is divided by 2 (since it is the second iteration) and added to the $simRatio$, in order to add less weightage to this value. After completing the iteration through the entire pathway, we compute the weighted sum of all these ratios to determine the $utilityScore$ for that pathway. Likewise, we apply these algorithms to calculate the $utilityScore$ and the $joinabilityRiskScore$ for all the pathways so that the users can compare among the different pathways available for a multi-way join of a specific cardinality. For computational efficiency, we have set a cutoff limit of 200 records for columns while calculating their similarity. Though this does not affect smaller datasets, for larger datasets, we can remove this constraint based on the availability of computational resources.

6.5.2 Join comparison view

LinkLens employs various visual cues to encode utility scores, joinability risk scores, and other relevant details for each pathway. In this sub-section, we discuss how different components of the Join Comparison View (Figure 6.4) assist users in

comparing multi-way join options and making an informed decision toward selecting a particular join pathway.

Encoding pathways: LinkLens represents join pathways using a series of connected rectangle boxes, where each box represents a dataset, and the connections denote possible joins. The size of each box reflects the relative size of the selected datasets in that pathway, based on the number of available records in those datasets.This provides the user with an estimate of the potential size of the joined results for each pathway. The rectangles are arranged in rows, with each row representing a different join pathway. The background color of the rectangles varies using a gradient from blue to white, indicating the order of the join in that pathway. For example, in Figure 6.4a, since datasets *Professional Medical Conduct Board A* and *Base Provider Enrollment File* are joined first, they have a blue background, while the subsequent join with dataset *Public Chauffeurs* results in a white background. If the pathway has a higher cardinality, such as ((D1 X D2) X D3) X D4, D1 and D2 would have a blue background, while D1, D2 and D3 will have a lighter blue background and the last dataset would have a white background. This visualization helps users compare different join cardinalities and understand the available pathways in those cardinalities **(T4)**.

Developing a join strategy requires understanding the connections between each dataset in a join. Hence, LinkLens displays the number of shared attributes using circles between the lines connecting datasets (Figure 6.4e). This allows users to quickly glance over the number of shared attributes between each join candidate and compare each pathway based on it joinability. On hovering over each circle, LinkLens shows the list of shared attributes (Figure 6.4d). Each dataset's record count, relative to others in the pathway, is displayed as a mini bar (Figure 6.4f). Hovering over these bars reveals the exact record numbers, giving users insight into whether the datasets are worth joining. These interactions supports human-in-the-loop exploration of

multi-way join pathways, allowing domain experts to use their background knowledge to select pathways with attributes relevant to their field. Attributes of interest selected earlier are highlighted in the same darker shade of grey, emphasizing their presence in the shared attributes and assisting in pathway selection based on user's interests. Upon selecting a pathway, LinkLens automatically suggests a join type, but users can augment this suggestion based on their needs, choosing from intersection join, master join, union join, or concatenation for each join action. Together, these visual analytic features in LinkLens assist the user in selecting a join strategy **(T6)**.

Encoding scores: LinkLens encodes the utility score and the joinability risk using two grey bars, with the filled area in each bar indicating the respective scores (Figure 6.4b, c). The utility score always falls within the range of [0, 100], establishing a part-to-whole relationship with the maximum possible utility score of 100 for any given pathway. Thus, the visual representation of the filled bar effectively illustrates this part-to-whole relationship, enabling users to quickly compare various pathways based on their utility scores **(T5)**.

While the pathways are arranged in descending order of utility scores, the joinability risk—represented by another filled bar—also aids in pathway comparison. Although these scores are visually abstracted through filled bars, hovering over them reveals a tooltip displaying the actual utility or joinability risk scores. It is crucial for users to select pathways with lower joinability risk, as no user wants to risk the disclosure of sensitive information. Yet, a pathway with higher utility may also carry some level of joinability risk. LinkLens helps users in balancing utility and privacy considerations in these multi-way joins, facilitating the selection of an appropriate join strategy **(T6)**.

6.6 Evaluate Join Outcome (G3)

Once the datasets are joined, users can inspect the join outcome to evaluate their utility (T7) and disclosure risk (T8). This is facilitated through Outcome Evaluation View in the LinkLens interface. In this section, we describe how various visual analytic interventions assist users in understanding the utility of the join outcome in relation to potential disclosures.

6.6.1 Methods for utility and disclosure evaluation

Finding useful attributes: The utility of the join outcome can be subjective, as it depends on the user's interests and may vary from one user to another. However, information-theoretic approaches, such as entropy, offer quantifiable measures in this context [198, 222, 199, 223]. This metric has also been used in analytical models that focus on comparing privacy and utility [224, 225]. Therefore, in LinkLens, we begin by calculating the entropy of the attributes of the join outcome dataset using the Shannon's entropy, as shown in Equation (6.3).

$$H(X) = -\sum_{i=1}^{n} P(x_i) \ln P(x_i) \tag{6.3}$$

where X represents an attribute in the join outcome, $H(X)$ denotes it entropy, and x_i represents each category of the attribute X in the join outcome.

The attributes are then arranged in descending order of their entropy. The idea behind this is that attributes with higher entropy indicate greater information content, which is beneficial to the user. This arrangement helps users prioritize attributes which may help them to understand the utility of the join outcome in a better way.

Finding disclosures: Attribute categories with very low frequencies can lead to disclosure of sensitive information. For example, if there is only one record for a person aged 11, this record can potentially expose sensitive details of that individual. Furthermore, if multiple quasi-identifiers are unique to this record, identifying

the individual becomes significantly easier. Therefore, addressing low-frequency attributes is crucial for identifying potential disclosures in the join outcome.

Thus, to identify potential disclosures, we analyze all unique combinations of the quasi-identifiers present in the join outcome. If any combination has fewer than or equal to k records, those records are flagged as possible disclosures. Ideally, k would equal 1; but we have noticed that some records become duplicated during the joining process due to null or missing values. One solution is to eliminate records with null values; however, this often leads to a significant loss of data quality, as many datasets contain numerous records with missing columns. These missing columns are often due to data quality issues or the redaction of sensitive information during publication. Thus, we have ensured that the values for the join keys are free of missing values; nevertheless, null values in other columns may still result in some duplicate records. As a result, we set $k = 3$ to detect disclosures in the join outcome. These disclosures can be evaluated through the Outcome Evaluation View of LinkLens.

6.6.2 Outcome evaluation view

This is the final screen of LinkLens, where the users can understand the dataset, make informed decisions, and export the dataset it as needed. In this sub-section, we describe how Outcome Evaluation View helps users in exploring the utility of the join outcome (T7) and inspect it for potential disclosures (T8).

Exploring join outcome: In LinkLens, we devised a visualization that organizes the attributes of the join outcome in descending order of their entropy. As previously noted, attributes with higher entropy signify greater information content; hence, prioritizing these attributes would be beneficial to users evaluating the utility of the join outcome. This visualization also includes a distribution representing the various categories for each attribute. The distribution is illustrated through a stacked bar chart, where each stack corresponds to a category of that attribute (Figure 6.5a).

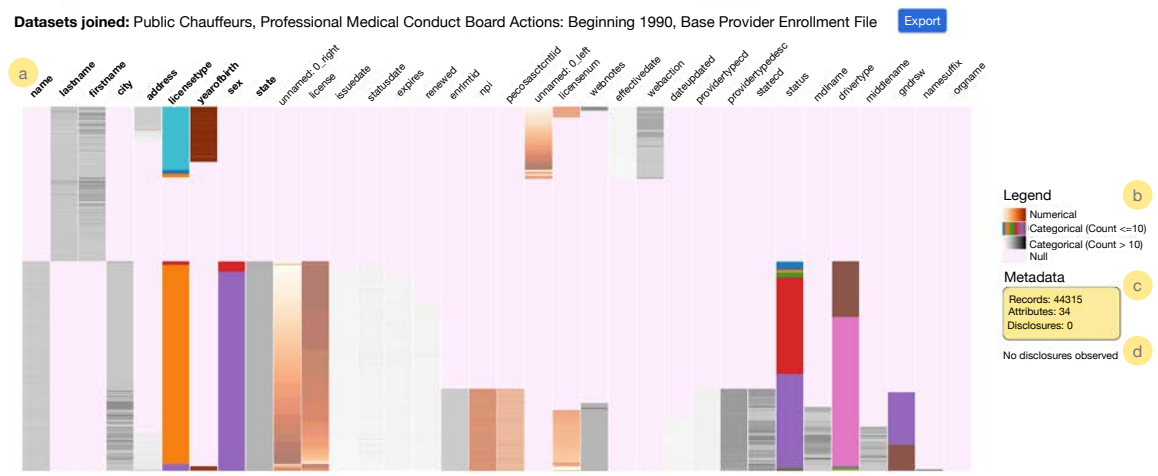


Figure 6.5 Outcome Evaluation View: Through this view, LinkLens assists user in (a) understanding the composition of the join outcome through a stacked bar visualization, (b) where different color schemes represent various attribute types and their values. (c) A metadata box summarizes some key components of the join outcome, while (d) showing any potential disclosures.

The categories are color-coded using different color schemes. Numerical attributes are represented through an orange sequential color scheme, with darker orange colors indicating higher numbers. Categorical attributes are represented using distinct colors from the Tableau 10 color scheme. If the number of categories are larger than ten, then each category's initial characters are converted to ASCII codes and assigned a corresponding color from a grey sequential color scale (Figure 6.5b). Null values, shown in light pink, provide insight into the completeness of the results and, consequently, the overall utility of the outcome dataset. For instance, in Figure 6.5, the join outcome has limited usefulness due to a high proportion of records containing null values. As our solution is interactive, users can explore record details associated with specific attribute categories by hovering over each category, which proves particularly useful for examining the outcome dataset at the record level. This colored categorization of the join outcome's records, along with the prioritization of high entropy attributes, helps users understand the outcome's composition and analyze it for utility **(T7)**. Users can also download the join outcome for further

investigation. A metadata box accompanying this visualization helps summarize the number of records, attributes, and disclosures (Figure 6.5c).

Acting on disclosures: The join outcome may or may not have any possible disclosures. Users can inspect them using the "Show Disclosure" button below the metadata box (Figure 6.5d; in this case, no disclosures are present). This reveals a small table below the visualization for the join outcome and populates it with the disclosure records. The attributes in this table view are arranged similar to that of the visualization of the join outcome which prioritize attributes with higher entropy. Users can inspect each disclosure record and take relevant actions for each of them **(T8)**.

In this table, each potential disclosure is loaded individually and includes two action buttons: one for keeping the record and the other for deleting it. Users can review the record and decide on an action based on the severity of the disclosure and their background knowledge. For instance, a user can opt to either remove the affected records or ignore them altogether. While other actions, such as redaction [226] or generalization [227, 228], can be performed for these records, they are beyond the scope of this work. Still, the user can identify those records through LinkLens and export them for further modifications. The exported version of the join outcome will contain all records except the ones the user wished to remove earlier.

6.7 Usage Scenario

Usage scenarios are essential for evaluating the practical applications of data analysis workflows. They demonstrate how users can interact with various features of a visual analytic tool to achieve specific goals and tasks, allowing for an assessment of the workflow's utility and usability in real-world contexts. In this section, we describe an usage scenario that help understand the efficacy of LinkLens in balancing utility and privacy factors during multi-way joins.

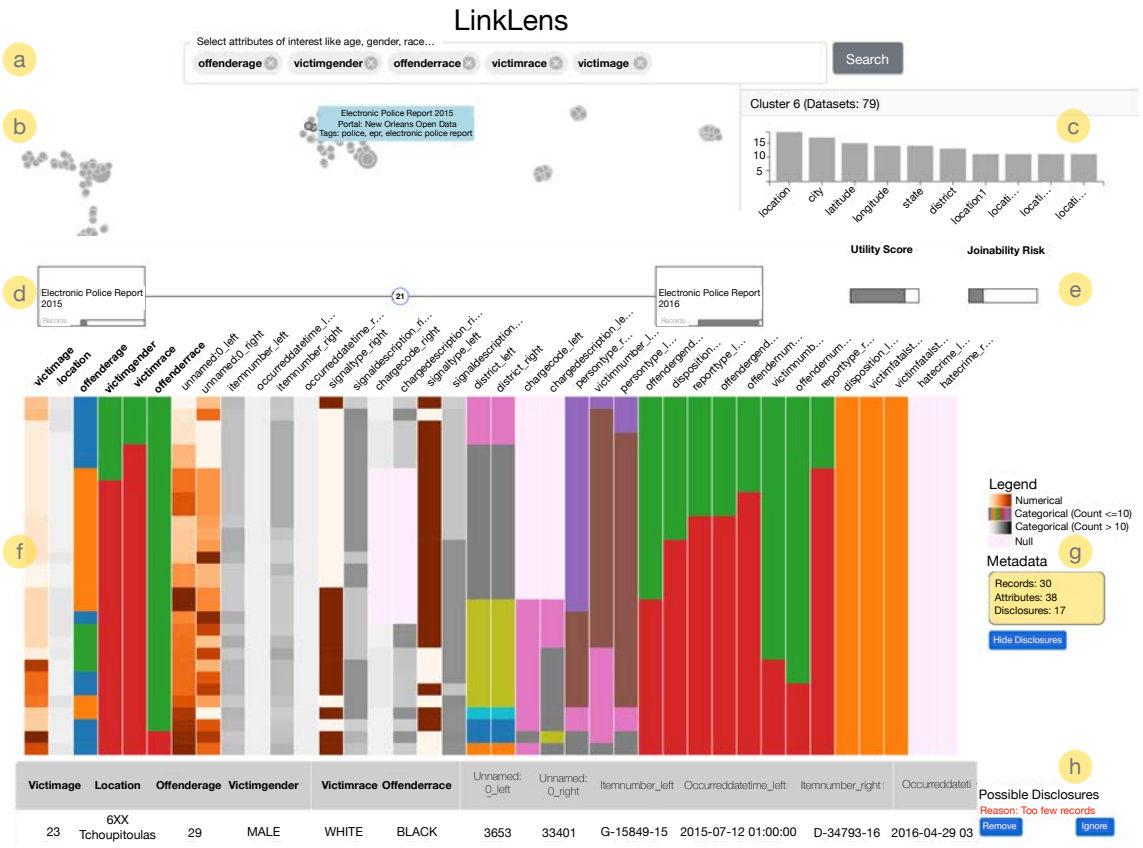


Figure 6.6 Usage Scenario:(a) A user can select different attributes of interest to search relevant datasets from the open data ecosystem. (b) LinkLens clusters them into joinable groups and (c) explains the rationale behind their grouping. (d) Users can then view the possible pathway(s) for the selected datasets and (e) compare their utility scores and joinability risks. (f) Next, users can select a join strategy and evaluate the join outcome by examining the distribution of the record categories. (g) A metadata box summarizes different characteristics of the join outcome along with the possible disclosures, (h) such as an incident where a data subject was charged with armed robbery.

Consider a scenario where Jane, a data analyst at a police department, is tasked with researching police incidents to train an AI model aimed at predicting future crime patterns and improving resource allocation. She wanted a tool that would allow her to join multiple open datasets but was also aware of the potential disclosure of sensitive information when linking datasets related to human subjects. Hence, she sought a tool that could balance privacy and utility factors during the data joining process. In this context, Jane opened LinkLens in her browser and selected a few attributes of interest from the available options, such as *offender*

age, *victim gender*, *offender race*, *victim race*, and *victim age*, commonly used in police datasets (Figure 6.6a) **(T2)**. Upon searching for relevant datasets, LinkLens clustered datasets from over 100 open data portals and displayed them in the Dataset Joinability View (Figure 6.6b). Jane used the Cluster Details feature to explore each cluster, examining the frequent attributes to understand the reasons behind the formation of that cluster (Figure 6.6c) **(T1)**. During her exploration, Jane discovered that Cluster 6 contained many location-related attributes. She investigated the datasets in this cluster and found two from the New Orleans Police Department [229]: *Electronic Police Report 2015* and *Electronic Police Report 2016* **(T3)**. Jane selected these two datasets and proceeded to the Join Comparison View to examine all possible combinations between them. The two datasets could only be joined by a single pathway (Figure 6.6d). In this view, LinkLens indicated that this pathway had both a high utility score of 79.75 and a significant joinability risk of 21 (Figure 6.6e) (**T4**, **T5**). Given the high utility of this pathway, LinkLens recommended performing an intersection join between these datasets **(T6)**. After joining them, Jane began exploring the utility of the join outcome in Outcome Evaluation View and gathered some insights, such as the majority of the victims being female (Figure 6.6f) **(T7)**. Although this was an important observation, it would not be useful for training her model, as the gender skewness could introduce bias. Additionally, she noticed that there were only 30 records in the join outcome, which was insufficient for her analysis and model training purposes (Figure 6.6g). Moreover, she observed some potential disclosures in the join outcome dataset. While exploring these 18 disclosure records, she observed an incident where a 23-year-old black male was charged with attempted robbery with a gun against a 29-year-old white male at 6XX Tchoupitoulas St on 12^{th} July 2015 at 01 : 00 hrs and again on 29^{th} April 2016 at 03 : 00 hrs with attempted simple robbery (Figure 6.6h) **(T8)**. At this point, Jane considered whether she could obtain a larger number of records for her research. To achieve this, she decided to

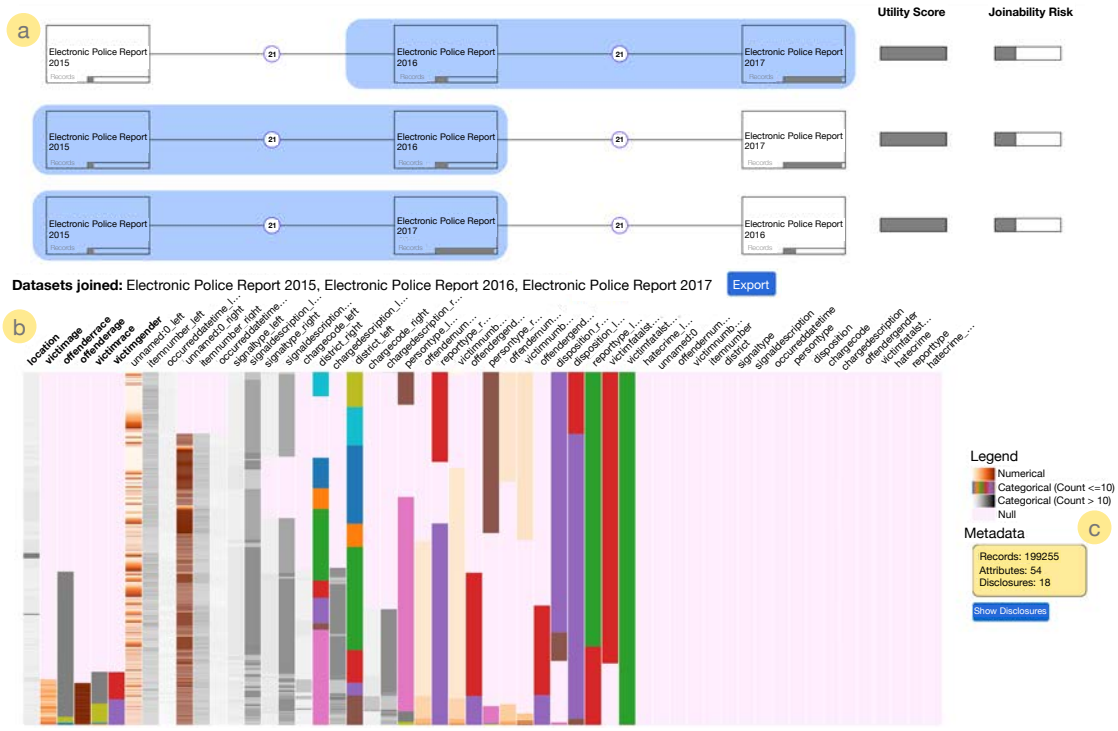


Figure 6.7 Usage Scenario continued: (a) If users select multiple datasets, LinkLens displays all possible pathways for this multi-way join and ranks them based on their utility score and joinability risk. (b) Users can explore the join outcome through the Outcome Evaluation View and (c) with the metadata on records, attributes, and potential disclosures, users can make a decision about which join outcome best suits their needs.

join the datasets using a union join. In a union join, records from both datasets are retained, and any common records are merged. This process yielded 262, 393 records, which Jane began exploring in Outcome Evaluation View. Although the balance between male and female victims improved with this join, a significant number of null records were present. Additionally, the potential disclosures increased to 74.

Jane wanted to check if she could increase the number of records before exploring the disclosures. She returned to Dataset Joinability View to search for a similar dataset that could provide additional records. After examining the datasets in the same cluster, she discovered the dataset *Electronic Police Report 2017* and selected it for a multi-way join. According to Equation (6.2), a combination of three datasets can produce three distinct pathways. LinkLens displayed all these potential pathways

in Join Comparison View and ranked them based on their utility scores in descending order (Figure 6.7a) **(T5)**. Although other pathways had marginally lower utility, the pathway *Electronic Police Report 2015* X (*Electronic Police Report 2016* X *Electronic Police Report 2017*) ranked the highest. Jane selected this pathway, and LinkLens once again recommended an intersection join due to the high utility associated with this pathway **(T6)**. Jane initially attempted this intersection join, but it did not yield any disclosures. Needing more records, she opted for a union join for this pathway, which produced 199, 255 records (Figure 6.7b). Although this was slightly fewer than the previous outcome, it included data for three years and reduced the possible disclosures to 18 (Figure 6.7c). At first glance, this join outcome seemed incomplete due to many null values. However, a closer look revealed that both join outcomes had 36 non-null or partially complete attributes, making them comparable on this metric. Jane realized that a round of data cleaning could resolve the null value issue, making the three-year join outcome more valuable for her training purposes **(T7)**. Additionally, this outcome was balanced in terms of gender, essential for training an AI model. She evaluated the disclosure records and decided to remove them since there were only about 18, a small number compared to the total records **(T8)**. Jane then exported this version for her research purposes. Thus, a researcher can use LinkLens to find relevant datasets, explore multi-way joining options, and balance privacy and utility factors throughout the process.

6.8 Conclusion

LinkLens represents a significant step forward in addressing the challenges of multi-way join analysis. This visual analytic system aims to help users navigate the complex task of balancing utility and privacy risks when joining multiple open datasets. By implementing attribute profiling based on user interests, as well as utility and joinability risk scoring algorithms, the LinkLens workflow provides a

structured approach to multi-way dataset join exploration and analysis. Interactive visualizations complement this by helping users discover joinable datasets from a vast pool, compare multi-way join options, and ultimately evaluate the join outcomes. The visual analytic interventions presented through Dataset Joinability View and Join Comparison View provide representations of dataset relationships and join strategies, aiding users in making informed decisions. Additionally, Outcome Evaluation View, with its entropy-based attribute prioritization and disclosure detection mechanisms, further supports the assessment of both utility and potential privacy risks in joined datasets. A key lesson learned from this work is the importance of design choices that help users effectively reduce the candidate space and triage options during the decision-making process. Furthermore, due to the computational complexity of joining large datasets, we also recognized the need for design and implementation choices that allow time for these computations while keeping the user informed throughout the process. By incorporating human expertise at crucial decision points while automating complex calculations, LinkLens aims to strike a balance between computational efficiency and domain-specific insights. This work contributes to the field of visual analytics and addresses important concerns in data privacy and utility maximization, potentially serving as a useful tool for researchers working with open datasets across various domains.

CHAPTER 7

FORTE: NET LOAD FORECASTING WORKFLOW

7.1 Introduction

The net load of an electric grid can be defined as the difference between the total electricity demand and the electricity generation from behind-the-meter resources such as solar and other distributed generators [230]. It can vary based on various factors, including weather conditions and the time of the day. Accurate net load forecasting enables grid operators, policymakers, and energy providers to make informed decisions regarding energy trade, load distribution, and resource allocation. However, the proliferation of solar energy generation sources in residential settings has significantly impacted the performance of traditional net load forecasting models [231]. We have collaborated with scientists who have developed a deep-learning model that produces probabilistic net load forecasts incorporating variables such as temperature, humidity, apparent power, and solar irradiance, achieving strong predictive performance and resilience in the face of missing data [61]. But, in order to improve trust in the model, these outputs need to be explored by domain experts, including scientists and grid operators. The model's performance may fluctuate due to seasonal variations in the input variables, and stakeholders must also assess its reliability in the face of noisy inputs mirroring real-life scenarios. Hence, the process is complex and time-consuming, prompting the need for an approach capable of performing these tasks and enhancing trust in the model's performance. Visual analytics can be instrumental here since prior research has shown that it can significantly enhance trust in model outputs during complex sense-making tasks [22]. In [23], it was argued that visual analytics would play a critical role in enabling

trust-augmented artificial intelligence and machine learning (AI/ML) applications in energy sector.

In light of this, we collaborated with energy scientists to thoroughly investigate the model's performance across diverse time periods and input scenarios, gaining valuable insights into the evaluation tasks needed to comprehend the model's effectiveness. Building upon this experience, we performed a design study to develop a system aimed at facilitating stakeholders in efficiently performing these evaluation tasks. As a result, we developed a visual analytics-based application `Forte` that empowers users to gain an in-depth understanding of the model's performance, effectively leveraging data visualization techniques to aid informed decision-making in the realm of energy planning and grid operations.

Our application aims to provide a broad understanding of various aspects related to net load forecasting. First, it enables researchers and scientists in the energy domain to assess net load variability concerning input variables by comparing model forecasts with actual net load values across different time periods and seasons. They can gain insights into their impact on model performance by analyzing the effects of variables like temperature, humidity, and apparent power on net load forecasts. Second, `Forte` helps evaluate forecast errors with noisy inputs at different noise levels, thus providing information for improving the model's reliability and robustness in real-world scenarios. This visual analytics-based approach can empower scientists and grid operators to make data-driven decisions, enhancing trust and confidence in the net load forecasting model.

While prior research has primarily concentrated on developing interfaces during the model development process, tailored to aid model developers, our focus lies in the post-hoc evaluation of the model's performance, catering specifically to the needs of energy scientists and grid operators [232, 233]. Other works explore the performance of probabilistic net load forecasting models through different visualization charts

but do not offer an integrated interactive interface [234, 235]. Our visual analytic tool, `Forte` presents an integrated workflow that empowers users to explore net load variability and forecast error analysis concerning various input variables and scenarios, which makes it a novel approach in this domain.

In this chapter, we first introduce our visual analytics-based application, `Forte`, developed in collaboration with energy scientists (Section 7.2). Emphasizing the analytical goals and tasks, we also provide insights into the underlying design rationale. Subsequently, we present observations gleaned through our application that can potentially drive advancements in grid operations (Section 7.3). Finally, we conclude with the lessons learned from this design study and how we incorporate them into our application (Section 7.4).

7.2 Visual Analytics-based Design

Our application `Forte` adopts a visual analytics-based design that integrates coordinated views aided with visual cues to show the various aspects of net load forecasting outcomes. It combines interactive visualization with performance metrics to instill greater trust in model outcomes, providing users with the flexibility to probe net load predictions as a function of input variables like temperature and humidity. In this section, we outline our `Forte`'s goals and tasks, followed by an explanation of our design rationale. We aim to achieve the following visual analytic tasks via `Forte`:

T1*: Understand actual net load and predictions across time periods and solar penetration levels:* The net load forecasting model is trained with data from varying solar penetration levels, while energy consumption fluctuates throughout the day and across seasons. Thus, this task involves comprehending the model's performance across diverse time spans and solar penetration levels.

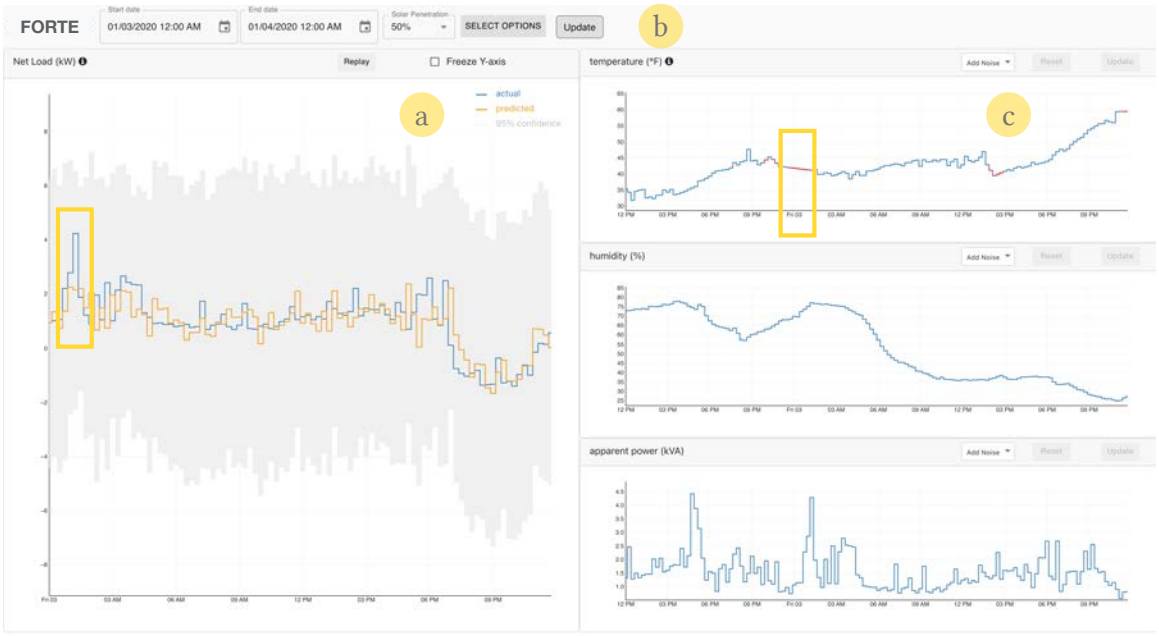


Figure 7.1 The interface for our net load forecasting visual analytic tool (`Forte`): (a) Our application facilitates the comparison of actual and predicted net load within the selected time frame and solar penetration levels as defined (b) through the Options Selection Area. Further, (c) the influence of various weather conditions on predictions can be explored via the Inputs View Area. The highlighted region shows instances of missing temperature data and resultant disagreement between predicted and actual net load within the same time period. These insights are valuable to grid operators as it allows them to review the data quality, evaluate its impact on model performance, and make recommendations for sensor/metering upgrade.

***T2**: Explore the impact of input variables on net load prediction:* Different inputs, such as weather conditions, influence the forecasting model differently. Therefore, it becomes essential to grasp the effects of these input variables on the model across diverse time spans, which motivates this task.

***T3**: Augment missing data with background knowledge:* Real-world weather data frequently includes gaps due to various factors. However, a domain expert may possess insights into expected temperature or humidity for specific timeframes. This task involves empowering users of the application to adjust inputs and observe how these modifications influence the model's performance.

***T4**: Design experiments simulating different noisy input scenarios:* Noisy inputs can vary due to factors like noise levels, direction (bidirectional/uni-directional), and number of observations. This task involves crafting scenarios using these factors to simulate and explore noise effects.

***T5**: Assess model efficacy across different months and varying levels of noise:* The model's sensitivity to noisy inputs can differ among months and noise levels. This task involves studying how varying noise levels affect model performance across different months.

Next, we explain our application's design by outlining its various high-level goals and demonstrating how it accomplishes these tasks.

7.2.1 Goal: understand net load forecasts w.r.t input variables

Understanding the interplay between net load forecasts and input variables is essential for making informed decisions in energy planning and ensuring efficient grid operations. Towards this end, `Forte` integrates three essential components:

Options Selection Area: As mentioned earlier, net load forecasts can fluctuate based on time periods and solar penetration levels. Accordingly, `Forte` offers these selections prominently at the top, within the Options Selection

Area (Figure 7.1b). This area contains two date and time pickers, enabling users to specify their preferred observation timeframe. Currently, users can opt for any date within the year 2020, with the potential for expansion as further data is available. Additionally, users can choose solar penetration levels from 0%, 20%, 30%, and 50%. This area provides options for choosing different prediction horizons (15 minutes or 24 hours ahead) and input variables (temperature, humidity, apparent power, etc.) tailored to user preferences. Initially, a limited set of input variables is loaded to reduce visual clutter, allowing users to add more based on their choices.

Net Load View Area: This component facilitates a direct comparison between the actual net load and the predicted net load for the chosen time period and solar penetration level, as selected within the Options Selection Area **(T1)**. This visual representation employs a blue line to depict the actual net load and an orange line to depict the predicted net load (Figure 7.1a). The extent of proximity/overlap between these lines indicates the level of agreement between actual and predicted net load, reflecting a superior model performance. But the degree of agreement can also be quantified by metrics like Mean Absolute Error (MAE) [236] and Mean Absolute Percentage Error (MAPE) [237], which are revealed by hovering on the icon button atop this area.

Since our net load forecasting model produces a probabilistic forecast, we additionally present the 95% confidence interval for this forecast, indicated by a subtle, shaded grey area. This design choice was made to streamline the view by avoiding the introduction of two additional lines, effectively minimizing visual clutter within this area. When users modify the options, we noted that the Y-axis within this area might shift due to value changes, impeding the observation of variations across distinct time periods or solar penetration levels. This problem can be alleviated using the "Freeze Y-axis" option, which, as the name suggests, freezes the Y-axis at the current values and plots the new values based on the frozen axis. Additionally,

changes can be tracked using the Replay button, which showcases net load changes through a slower animation ($\approx$ 10s).

Inputs View Area: Located on the right-hand side of the application, the Inputs View Area displays the selected inputs (as selected from the Options Selection Area) and their respective values during the chosen time period (Figure 7.1c). It also shows some of the historical data used while generating the forecast for this period. This visualization aids in establishing correlations between weather data and the agreement/disagreement observed between the actual and predicted net load, thereby impacting model performance **(T2)**.

Nonetheless, weather data might feature gaps for specific time spans, which are addressed through linear interpolation connecting the nearest available data points. These interpolated points are indicated in red, and users have the flexibility to drag and adjust them based on their expertise **(T3)**. The data quality, denoting the percentage of missing data, can be accessed by hovering over the icon button atop each input variable. Conversely, if the users do not trust the quality of the available weather data, they can apply a uniform noise of 5% or 10% via the "Add Noise" option for each input variable. All these changes are reflected on the Net Load View Area once users hit the "Update" button. Thus, `Forte` begins with simple visualization and default settings, easing the learning curve as users delve into advanced features.

7.2.2 Goal: compare model performance w.r.t noisy inputs

During our investigation into the influence of input variables on net load prediction, we noticed that the model's responses varied with different noise levels. Hence, we developed a separate linked page with the following components.

Experiment Design Area: Users can generate simulated noisy inputs for varying dates spanning multiple months and a specific input variable. This area empowers users to select their preferred input variable (temperature, humidity,

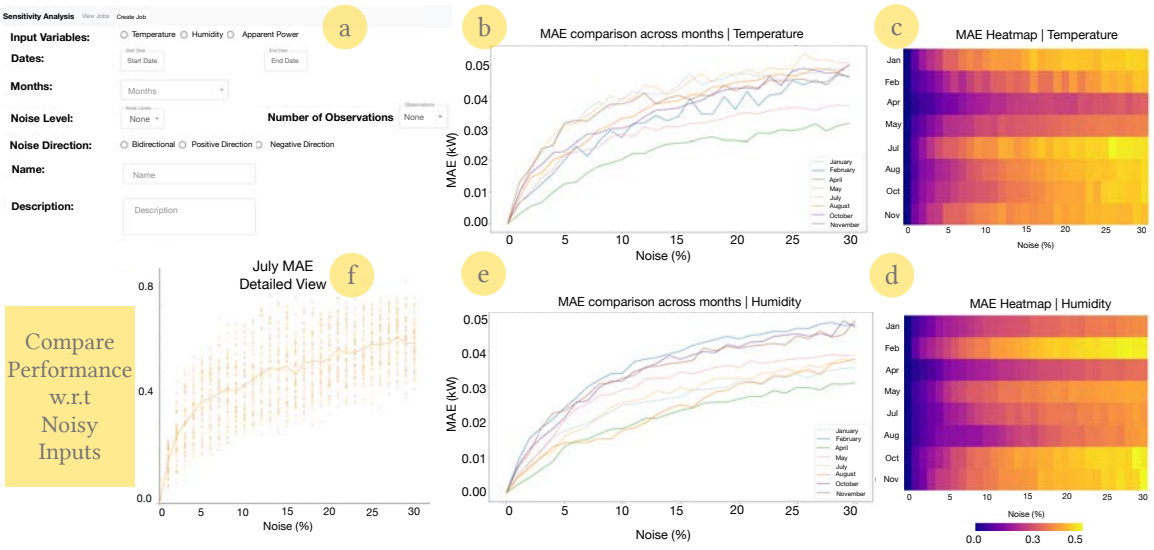


Figure 7.2 Experimental Results: (a) Our application `Forte` enables the design of experiments through the creation of noisy inputs using various factors; and the results (error rates) can be cross-compared across various months for both the input variables of (b, c) temperature and (d, e) humidity; (f) with the option to view detailed observations for each month. These insights generated through `Forte` are valuable to the user (a grid operator) to not only reveal the underlying dependence of the model outcome (net load prediction) on different input weather conditions but also better prepare ahead of any impending weather events (e.g., heat/cold wave).

apparent power), set start and end dates, and designate desired months for introducing noise **(T4)**. The area also offers the flexibility to add or subtract a uniform noise (ranging from 1% to 30%) from the original inputs or use a combination thereof (Figure 7.2a). As the noise is uniformly distributed, the Experiment Design Area accommodates multiple observations/forecasts with identical inputs. Finally, users can add a name and short description for future reference, and our application will show an estimated time for completing this experiment. This experiment-based architecture enables `Forte` to manage computational overload efficiently.

Experiments View Area: Once the designed experiments are complete, those are available for the user's perusal. Users can select any of the completed experiments/jobs from the left-hand side navigation bar. Each experiment initially displays two line charts depicting the error metrics MAE and MAPE (Figure 7.2b, 7.2e). These charts comprise lines corresponding to the months chosen during the experiment

design. Each line illustrates the deviation in error metrics from their baseline values (established at 0% noise or no noise) **(T5)**. We offer an alternative visualization in the form of a heatmap, illustrating the deviation in MAE from the baseline values for each month. Based on initial feedback, users found this heatmap particularly useful for comparing the model's sensitivity across different months (Figure 7.2c, 7.2d). In addition to this, users may want to explore the error rates for each month. Hence, we also include two scatterplots for each month (for each of the error metrics), which show the error rates for each of the observations (Figure 7.2f). This scatterplot is then augmented with a line showing the average error rate for that month across different noise levels, mimicking the corresponding line in the first line chart. This view area helps to understand the model's performance when faced with noisy inputs and, in the process, improves trust in the model. `Forte` is primarily developed using React.js and D3.js for the frontend, and Flask framework in Python for the backend.

7.3 Experimental Results

In this section, we showcase some outcomes from our application, emphasizing their possibility to enhance model training and potentially streamline grid efficiency. We illustrate this through a practical scenario involving a research scientist named Amy. Amy seeks to comprehend the influence of noisy inputs on the model and enhance trust in net load forecasts, consequently aiding effective grid operations planning.

Amy scrutinized net load predictions for January 3rd to 4th, 2020, at a 50% solar penetration level through our application **(T1)**. She noted a general alignment between predictions and the actual net load values, barring a deviation around 1 a.m. on Friday, January 3rd (Figure 7.1). Intrigued by this discrepancy, she used our application to delve into the temperature data for that period **(T2)**. Here, Amy observed some missing data around the same time. The data around this timeframe underwent linear interpolation using the nearest available data, which Amy speculated

might contribute to the discrepancy. To investigate, she interactively adjusted the line chart at this point, thereby updating the temperature values for that period **(T3)**. This led to slight prediction variations, suggesting temperature's significance as an input variable to the model. As a further step, she introduced a uniform 5% noise to all temperature values within the selected time period. Interestingly, this led to several changes in the predictions.

Now, Amy aimed to systematically comprehend the influence of noisy temperature values on the net load forecasting model. She devised an experiment by (arbitrarily) selecting the 3rd and the 4th days of various months in 2020 and introducing consistent bias/noise, ranging from 1% to 30%, to the recorded temperature values **(T4)** (Figure 7.2a). Surprisingly, she observed no change in the error metrics. Her inference was that the model normalizes inputs before generating the predictions, thus explaining the similar outputs despite varying noisy inputs. Subsequently, Amy replicated the experiment, introducing uniform noise to the temperature values. As an example, for a temperature of 60°F with 10% added noise, the range of noisy input could span from 60°F to 66°F. Given the randomized nature of this experiment, she chose to replicate it across 50 observations/iterations, akin to repeated measures design [238, 239]. Having reviewed the findings of this experiment, she proceeded to repeat it over eight months (January, February, April, May, July, August, October, and November). Her observations unveiled that, although error rates were minimal, the model displayed heightened sensitivity to noisy data during January and July, across numerous noise levels—albeit with exceptions. In contrast, the model exhibited the least sensitivity during April and May (Figure 7.2b and 7.2c). The observed variations in the model's sensitivity to noisy perturbations in temperature data across different months, can be attributed to the influence of seasonal weather variations on usage of electricity **(T5)**. For example, typically the heating and cooling load – which drives the residential energy

demand – typically peaks during the coldest (e.g., January) and the hottest (e.g., July) months, thereby ensuring heightened sensitivity of net load to temperature variations. In contrast, sensitivity of residential energy usage to temperature perturbations remain low in shoulder months (e.g., April and May) with typically milder weather. Given the potential impact of climate change on these results over time, it is imperative to use `Forte` to conduct further such experiments regularly. Insights revealed from this experiment helped Amy gain confidence in the model. This also underscores the embedded learning process, as these insights serve as valuable resources for retraining the model to handle noisy scenarios better.

Subsequently, Amy sought to determine if humidity yielded similar effects on the model's performance. She initiated a parallel experiment focusing on humidity (Figure 7.2d and 7.2e). Notably, her observations indicated heightened model sensitivity during February and October, with reduced sensitivity aligning with April—mirroring the earlier temperature findings. Furthermore, she delved deeper into the predictions for each month, aiming to grasp the distribution of error metrics across noise levels within 50 observations (Figure 7.2f). While noting the presence of outliers in these error metrics, Amy observed that the mean line of these observations effectively captured the trend across most months. On an overall assessment, Amy discerned that while error rates varied across different months, the model consistently demonstrated commendable performance, with notably low error rates difference from the baseline (≈ 0.05 kW MAE). We can thus conclude that our visual analytics tool `Forte` effectively enabled her to grasp the model's performance concerning diverse weather data and their noisy variants, thus improving her trust in this model.

7.4 Conclusion

The significance of accurate net load forecasting in energy planning and grid operations cannot be overstated. Hence, in this study, we explored net load

forecasting, leveraging a collaborative approach with domain experts to develop a visual analytics-based application. By partnering with energy scientists, we not only identified critical evaluation tasks but also translated them into an intuitive interface that empowers users to make sense of complex model behaviors.

Throughout this endeavor, we gained valuable insights into the challenges posed by noisy inputs, seasonal variations, and the need to instill trust in forecasting models. The collaborative process exposed the complexities of real-world data analysis and emphasized the necessity for efficient, user-friendly tools that bridge the gap between model insights and actionable decisions. We can argue that our application is a first step towards this direction. As a next step, we plan to elucidate the input normalization process, add other error metrics, and incorporate economic planning and analysis to enable stakeholders to gauge the cost-benefit ratio and enhance trust in the net load forecasting model [240, 241].

Looking ahead, there are multiple areas where we would like to enhance `Forte`. Incorporating more datasets spanning multiple years, expanding to additional weather conditions, automating aspects of the experiment design area, and even using transfer learning techniques are all promising avenues for future exploration. As the energy landscape evolves, our application's adaptability will play a vital role in helping energy planners, grid operators, and policymakers navigate the complexities of net load forecasting. In conclusion, our collaborative effort has yielded a powerful tool that not only deepens our understanding of net load predictions but also lays the foundation for more informed and efficient energy planning decisions in the years to come.

Acknowledgment

Parts of this work were supported by the U.S. Department of Energy Solar Energy Technologies Office and the Office of Electricity Sensors Program, and

performed jointly at the Pacific Northwest National Laboratory under Contract DE-AC05-76RL01830 and at the Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

CHAPTER 8

WORKFLOW FOR TRUST-AUGMENTED MODEL COMPARISON

8.1 Introduction

Comparing multiple computational models for their performance is crucial for enhancing trust in model outcomes because it provides a basis for evaluating the reliability and consistency of each model [242, 243, 244]. By assessing how different models perform under various conditions and scenarios, stakeholders can better understand their strengths, weaknesses, and overall effectiveness. Such comparative analysis helps to identify the most suitable model for specific tasks or applications, thereby instilling confidence in the reliability of the chosen model. This becomes more important while predicting the net load of an electric grid, which is defined as the difference between total electricity demand and generation from behind-the-meter resources like solar and distributed generators, and is influenced by various factors such as weather conditions and time of day [245, 230].

Accurate net load forecasting enables grid operators, policymakers, and energy providers to make informed decisions regarding energy trade, load distribution, and resource allocation. However, the rise of solar energy generation sources in residential settings has significantly impacted the performance of traditional net load forecasting models, highlighting the need for robust time-series forecasting techniques [231]. Collaborating with scientists, we developed a deep-learning model that integrates variables such as temperature, humidity, apparent power, and solar irradiance to achieve strong predictive performance and resilience in the face of missing data [61]. Still, the model's sensitivity to seasonal variations and noisy inputs underscores the need for further exploration with domain experts. Hence, we developed an interactive

tool that empowers users to investigate the model's performance across diverse time periods and input scenarios [28].

While we initially recognized the importance of comparing the model's performance with traditional models, feedback from domain experts further emphasized its significance in building trust in model outcomes. Visual analytics can play a pivotal role here, as evidenced by prior research demonstrating its importance in enhancing trust in machine learning models [246]. This is exemplified by the interactive tool we developed, which enabled domain experts to extract valuable insights concerning the model's sensitivity toward temperature and humidity. Moreover, recent discourse, as highlighted in [23], emphasizes the critical role of visual analytics in fostering trust-augmented applications of artificial intelligence and machine learning (AI/ML) within the energy sector. Building upon this, we designed our application, incorporating carefully selected visual analytic interventions. These interventions facilitate the comparison of multiple models across various parameters, including solar penetration levels, dataset resolutions, and different hours of the day, enhancing stakeholders' confidence in model performance.

The aim of our application is to build trust through model comparison, primarily comparing our net load forecasting model with a reference model. We enhanced our model to process inputs at varying resolutions and then devised the reference model, which generates predictions by averaging net load ground truths for the last 30 days at the same time point. Despite lacking predictive ML components, this reference model serves as a benchmark in various net load forecasting competitions, including the recent Net Load Forecasting Prize by the National Renewable Energy Laboratory (NREL) and the U.S. Department of Energy Solar Technologies Office (SETO) [247]. Through the web interface of our application, we were able to uncover patterns in the performance that help improve trust in the model outcomes. In this work, we

identify the visual analytic tasks that are required to compare these models. These tasks can be extended to compare other similar net load forecasting models, enabling users to make well-informed decisions based on the outcomes of these models.

While previous research has predominantly focused on developing interfaces during the model development phase, tailored to assist model developers, our emphasis lies in the post-hoc evaluation of model performance, specifically addressing the needs of energy scientists and grid operators [232, 233]. Other studies have explored the performance of probabilistic net load forecasting models through different visualization charts but have often lacked an integrated interactive interface [234, 235]. In contrast, our interactive visual analytic tool provides carefully designed visual cues for comparing model performance across different factors like dataset resolutions and solar penetration levels, thus offering a novel approach in this domain.

In this chapter, we first introduce the different models used in this work and the rationale behind choosing them. Subsequently, we outline the identified visual analytic tasks and detail the design decisions guiding the development of our interactive interface. This is followed by some of the observations made by our power scientist collaborator through our application that demonstrate its efficacy in comparing model performance across multiple facets. Finally, we conclude by sharing insights gained from this development and discussing some of the future research opportunities in this domain. Additionally, a short demonstration video is available here.

8.2 Model description

We start with a deep learning-based probabilistic model tailored for net load forecasting in high behind-the-meter solar scenarios [61]. This model has three key components: a kernelized probabilistic forecasting (kPF) module, an autoencoder

(AE), and a long short-term memory (LSTM) network. This model effectively captured complex temporal dependencies and uncertainties inherent in net load data, which is crucial for reliable forecasting in environments with high solar penetration. By incorporating kernel methods into probabilistic forecasting, the model handled non-linear relationships and captured subtle variations in net load influenced by solar energy fluctuations. The autoencoder component enhanced feature extraction and dimensionality reduction, facilitating the LSTM network's ability to capture long-term dependencies and predict future net load values accurately. Experimental results demonstrated the superior performance of this model compared to traditional forecasting models, showcasing its efficacy in addressing the challenges posed by high solar scenarios and advancing the state-of-the-art in net-load forecasting methodologies.

This model was used in the Net Load Forecasting Prize competition hosted by NREL and SETO [247]. However, during this competition, we observed its underperformance on the data provided by the organizers, which had lower resolutions. We discovered that while the autoencoder component excelled with high-resolution datasets (such as 15-minute intervals), it struggled to effectively capture temporal dependencies in lower-resolution datasets (e.g., 1-hour intervals provided by the organizers). Consequently, this limitation led to subpar outcomes generated by the LSTM component. In light of this, we opted to remove the kPF and autoencoder components and instead developed a version of the model solely utilizing the LSTM component. Additionally, fine-tuning the number of layers in the LSTM component yielded significantly improved results in the competition.

The competition employed a reference model to assess model performance across various probability levels. As previously mentioned, this reference model simply utilizes historical input data from the past 30 days to generate probabilistic forecasts for a specific time point. This model can serve as the initial benchmark for

assessing the effectiveness of other models. Therefore, in this work, we developed a reference model following the same principles. Since these forecasts are probabilistic in nature, we calculated the Continuous Ranked Probability Score (CRPS) for both the reference model and our model. Subsequently, we computed the Continuous Ranked Probability Skill Score (CRPSS) based on these CRPS scores, evaluating whether our forecast presents an improvement or deterioration compared to the reference forecast [248, 249]. A positive CRPSS indicates that the forecast outperforms the reference forecast, whereas a negative value suggests inferior performance. Utilizing these CRPSS values, we compare the performance of our model against the reference model across multiple dates throughout the year, and present these values in our application.

8.3 Visual Analytics-based Design

Our application employs a visual analytics-based design featuring coordinated views enhanced with visual cues to help users compare model performance. It combines interactive visualization with comparison metrics like CRPSS to improve trust in the model outcomes and also allows the users to probe the model and understand its performance across different solar penetration levels and months. In this section, we highlight the two tasks performed by our application and how its visual analytics-based design aids in executing these tasks:

T1*: Compare model performance across different solar penetration levels and data resolutions:* Model performance may vary across different solar penetration levels due to the increased variability in net load data caused by intermittent solar generation. LSTM-based models might struggle to capture and predict these dynamic behaviors accurately. Furthermore, datasets with higher resolutions, such as sub-hourly intervals, enable models to more effectively capture short-term fluctuations and

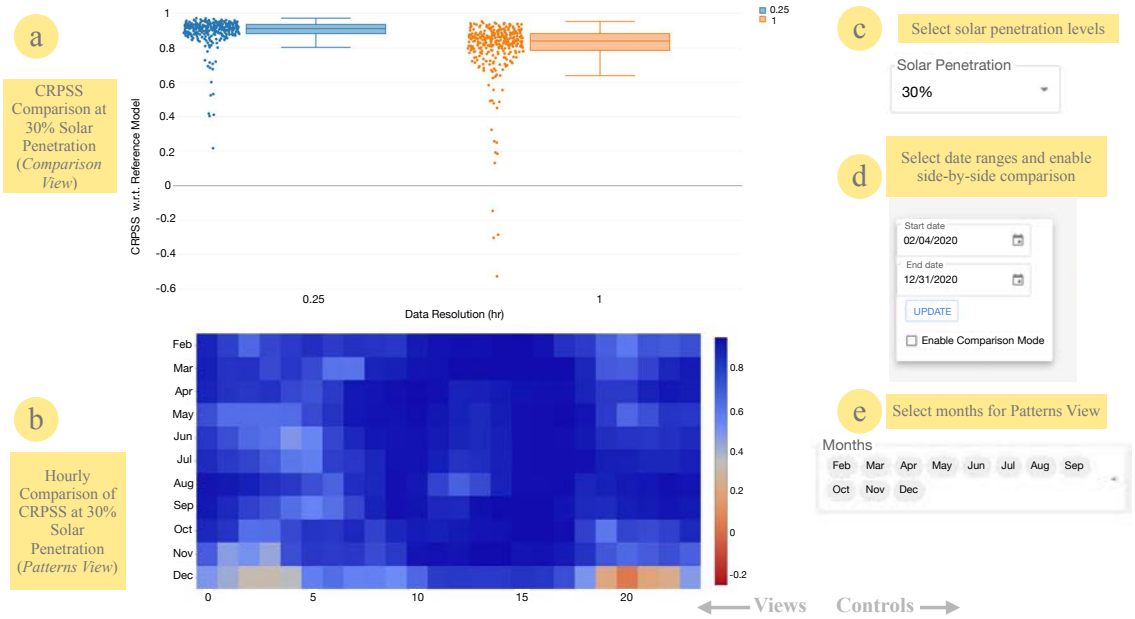


Figure 8.1 Visual analytic application: (a) The Comparison View the facilitates comparison of CRPSS values between the net load forecasting model and the reference model at various data resolutions throughout the year. (b) The Patterns View aids in identifying performance trends across different hours of the day and months. (c), (d) and (e) denote filters for selecting different solar penetration levels, start and end dates, and specific months for the heatmap, respectively.

dependencies. Hence, this task essentially involves comparing model performance at different solar penetration levels and data resolutions.

***T2**: Identify patterns across different timeframes:* By analyzing performance across multiple timeframes, power scientists can assess the net load forecasting models' robustness and consistency in capturing both short-term fluctuations and long-term trends. This evaluation helps to identify whether a model's performance is consistent across various temporal scales or exhibits variability or biases at specific time periods. This task relates to identifying patterns in the model's performance across different months, various hours of the day, and different time periods.

Our application implements multiple coordinated views and components in order to fulfill these tasks. Next, we discuss the design of these views and components along with the rationale behind them:

***Comparison View**:* As comparing model performance is the main objective of our application, we begin with the Comparison View, which utilizes modified box plots to compare models. Figure 8.1a depicts CRPSS values on the y-axis and different data resolutions on the x-axis. The box plots illustrate the distribution of CRPSS values, with the median typically exceeding zero, indicating superior performance of our model over the reference in most cases. Users can utilize the solar penetration level filter to compare performance across various levels (20%, 30%, 50%) (Figure 8.1c) **(T1)**. However, based on initial feedback from domain experts, we recognized the importance of displaying the distribution of CRPSS values. Therefore, we integrated dots representing the CRPSS values, jittered along the x-axis to illustrate their distribution over the entire year. These dots align with insights from the box plot and additionally reveal instances where our model performs worse than the reference on certain dates. This led to the development of the Patterns View, where we can identify the timeframes where the model underperforms compared to the reference.

Patterns View: In this view, our application utilizes a heatmap to depict performance patterns for each month across different hours of the day **(T2)**. The x-axis represents the 24 hours of the day (0-23), while the y-axis displays the months of the year (Feb - Dec) (Figure 8.1b). Each box in the heatmap denotes the average CRPSS value for each month at each hour, indicated by the color. Darker blues signify more positive CRPSS values, indicating the superior performance of our model compared to the reference at that time. Conversely, darker reds indicate more negative CRPSS values, signifying poorer performance of our model compared to the reference at that time. This diverging color scale aids in easily identifying performance patterns across different time points and facilitates the identification of instances where our model does not outperform the reference. Users can filter specific months to focus on particular time periods and analyze performance accordingly (Figure 8.1e). Additionally, based on feedback from domain experts, we implemented the Sidebar component to allow users to select specific date ranges and analyze performance patterns within those ranges.

Sidebar: The Sidebar facilitates user selection of start and end dates to filter results across all views, enabling focus on specific date ranges **(T2)** for comparing model performance within those periods (Figure 8.1d). Additionally, based on feedback from domain experts expressing interest in comparing model performance across all solar penetration levels simultaneously, our application offers a comparison mode toggle in the Sidebar. Enabling this mode updates both views to display box plots and heatmaps for all solar penetration levels side by side, aiding in point-to-point comparison and identification of performance patterns across different solar penetration levels.

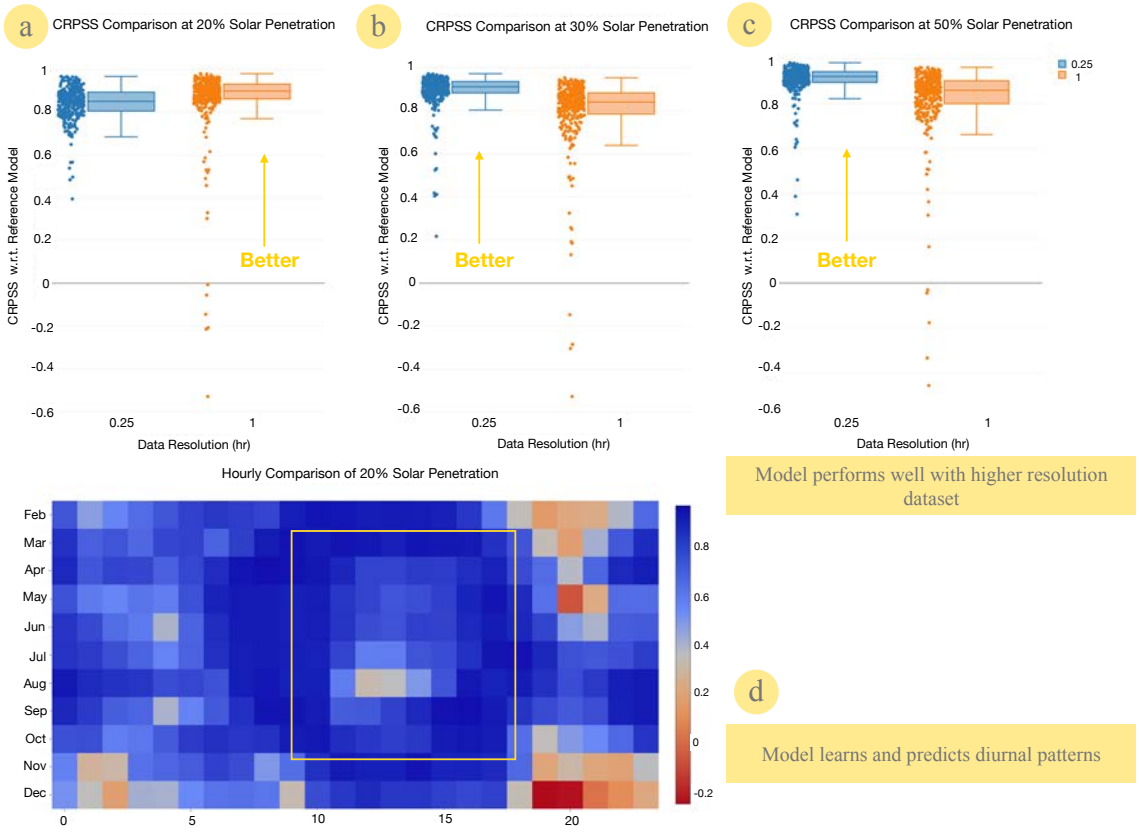


Figure 8.2 Results from a case study: (a), (b), (c) display CRPSS values at varying solar penetration levels, highlighting the model's superior performance with higher-resolution datasets. (d) Additionally, our application reveals insights such as the model's ability to learn and predict diurnal patterns, as evidenced by highlighted box-like patterns.

8.4 Results From A Case Study

The efficacy of an application can be validated if the application is able to perform the intended tasks effectively. In this section, we show the results through a case study that demonstrates how our application can be used to compare and select net load forecasting models effectively.

This case study involved a power scientist with over 10 years of experience in power and grid systems. With expertise in nonlinear dynamics, large-scale networks, and distributed control, he played a crucial role in developing the model. His main objective was to assess the model's performance relative to the reference model across various time points and solar penetration levels, which are critical

factors to consider before deploying it for a project. We informed him that we had integrated CRPSS values for both the model and the reference model for all dates throughout the year. He then accessed our application through a browser and examined the distribution of CRPSS values across different data frequencies at a 20% solar penetration level (Figure 8.2a). Notably, he discovered that the model performed better with the lower resolution dataset (1-hour), contrary to expectations for an LSTM-based model **(T1)**. Upon further examination, he noted a marginal difference in the median CRPSS between high and low-resolution datasets (0.85 and 0.89, respectively). Therefore, he enabled the comparison mode through the Sidebar, enabling him to assess the model's performance as solar penetration increased. He noted that the model consistently performed well with the higher resolution dataset (15-min) across all other solar penetration levels, confirming the notion that our model excels with high-resolution datasets in high solar penetration scenarios (Figures 8.2b and 8.2c). Next, he observed that although the median CRPSS was significantly above zero, the minimum value was negative. Upon inspecting the dots adjacent to the box plots in the Comparison View, he observed that while most dots clustered around the median line, there were a few outliers with negative CRPSS values.

Hence, the scientist sought to understand temporal patterns to identify if any specific hour of the day contributed to the negative values. Consequently, he navigated to the Patterns View within our application to analyze the CRPSS value distribution in the heatmap at a 20% solar penetration level (Figure 8.2d). Consistent with observations from the Comparison View, most heatmap boxes displayed varying shades of blue, indicating superior model performance compared to the reference across most months and hours. On closer look, the scientist noted a box-like pattern in the heatmap, revealing enhanced performance during morning hours (8 am to 10 am) from April to September, followed by a decline during midday (11 am to 4 pm), and another spike in the evening (4 pm to 6 pm) during these months. This

pattern suggested that the model effectively captured diurnal variations in net load data and adjusted predictions accordingly **(T2)**. While similar box-like patterns were evident across other solar penetration levels, the intensity of model performance varied with increased solar penetration. This insight holds practical significance for model selection during deployment, as it suggests the potential use of different models or model ensembles tailored to distinct times of the day, leveraging their respective performances on diurnal patterns. Thus, the power scientist was satisfied that our application could yield valuable insights regarding the model, aiding informed decision-making.

8.5 Conclusion

Efficient model selection for net load forecasting plays a pivotal role in energy planning and grid operations. In this work, we delve into this process, integrating visual analytics with input from domain experts to identify key tasks for comparative model selection. Subsequently, we translate these tasks into an interactive interface, enabling users to assess model behavior across various factors such as solar penetration levels, data resolution, and time of day.

Throughout this endeavor, we gained valuable insights into the model behavior and the challenges posed during the multi-way comparison of models. This collaborative effort underscored the need for interactive tools that facilitate the seamless translation of model insights into actionable decisions. We can argue that our application is a first step towards this direction. As a next step, our plan is to incorporate multiple net load forecasting models into the application and integrate additional metrics for effective comparison of their performance.

Looking ahead, we aim to enhance our application in several ways. Incorporating economic planning and analysis options will allow stakeholders to assess the cost-benefit ratio before model selection, thereby enhancing trust in the

outcomes. Additionally, as the energy landscape evolves, our application's flexibility will be pivotal in effectively comparing and selecting forecasting models for real-world applications. In summary, our collaborative endeavor has produced a robust tool for multi-faceted model comparison and has paved the way for informed decision-making through visual analytics in energy planning.

Acknowledgment

Parts of this work were supported by the U.S. Department of Energy Solar Energy Technologies Office and the Office of Electricity Sensors Program, and performed jointly at the Pacific Northwest National Laboratory under Contract DE-AC05-76RL01830 and at the Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344).

CHAPTER 9

CONCLUSION

This dissertation has explored the critical intersection of privacy preservation, data visualization, and analytical uncertainty mitigation in the context of open data ecosystems and net load forecasting. Through a series of interconnected studies and system developments, we have addressed the pressing need for robust privacy-preserving techniques in data visualization and analysis, while also tackling the challenges of analytical uncertainty in complex domains such as energy forecasting and disclosure management. The work undertaken in this dissertation aims to establish a foundation through interactive workflows and tools that help transform current "unknown unknowns"—such as the risk of disclosure in open data—into more manageable "known unknowns." Additionally, it evaluates whether our workflows and tools can effectively handle challenges that already fall into the "known unknown" category, offering valuable insights for future advancements in these areas. By pushing these boundaries, this dissertation lays the groundwork for better-informed decision-making in privacy-sensitive and data-intensive applications. In this chapter, we will provide a summary of the work presented throughout this dissertation, highlighting key contributions made across the various chapters. We will conclude by exploring the potential future directions and broader implications of this research, outlining opportunities for continued development in privacy-preserving data visualization and analytical uncertainty management.

Review of the literature landscape: Our journey began with a survey of privacy-preserving data visualization techniques. This review revealed a landscape where visualization plays a crucial role in empowering various stakeholders in the data ecosystem to understand and manage privacy implications. We identified key tasks

such as hiding data, evaluating risks, understanding policies, evaluating trade-offs, and comparing algorithms. The survey also highlighted significant research gaps, including the need for uncertainty visualization in privacy contexts, dynamic risk visualization, and privacy-aware citizen science. This foundational work set the stage for our subsequent investigations and system developments.

Vulnerable datasets discovery: Building on the insights from our literature review, we conducted a red-teaming exercise to identify vulnerabilities in open datasets. Collaborating with data privacy experts, we identified several attack scenarios and compiled a list of vulnerable datasets from over 100 open data portals. This ethical hacking approach also revealed several concerning examples of how seemingly innocuous open data could be combined to disclose sensitive information about individuals. Our findings underscored the urgent need for proactive risk assessment tools in the open data ecosystem. This chapter's work directly informed the development of our subsequent risk inspection workflow, PRIVEE, and highlighted the importance of considering both individual datasets and their potential combinations when assessing privacy risks.

Disclosure inspection workflow: Responding to these vulnerabilities identified in Chapter 3, we developed PRIVEE, a visual analytic workflow for disclosure risk inspection in open datasets. This system empowers data defenders to triage joinable groups of datasets, compare joinability risks, and identify specific cases of disclosure. We distill this workflow through a web-based interface using React.js and d3.js for the front end and Python for the backend API. PRIVEE's design incorporates interactive visualizations that provide transparent explanations of risk assessments, allowing defenders to make informed decisions about data release and privacy protection strategies. The workflow's effectiveness was demonstrated through case studies with domain experts, showcasing its potential to significantly enhance privacy protection in open data ecosystems.

Utility calibration workflow: After developing the disclosure inspection workflow, we shifted our focus to calibrating the utility of linked datasets. We recognized that this aspect deserved to be a stand-alone workflow, enabling the evaluation of dataset joins independently of other privacy considerations. To address this need, we developed VALUE, a system specifically designed to assess the utility of joining open datasets. This work is built upon the concepts introduced in PRIVEE, expanding the scope to allow users to evaluate the potential benefits of joining datasets. VALUE's interface facilitates the exploration and comparison of various join combinations across multiple open data portals, offering a framework for decision-making in data sharing and analysis. This chapter emphasizes the importance of evaluating utility factors when joining open datasets, ensuring that informed actions are taken in the face of analytical uncertainty, particularly when determining which open data best meets the user's needs.

Balancing privacy and utility factors in multi-way joins: Building on the foundations laid by PRIVEE and VALUE, we developed LinkLens to address the more complex challenge of multi-way joins. This chapter introduces a workflow that balances privacy considerations with utility factors, helping researchers navigate analytical uncertainty when combining datasets from various open data portals, which could otherwise result in a combinatorial explosion. LinkLens represents a significant advancement in our ability to manage privacy risks in increasingly complex data environments while still extracting valuable insights from diverse data sources. The system's interface enables users to explore and compare different join combinations across multiple open data portals, providing a framework for decision-making in high-consequence scenarios. Furthermore, LinkLens incorporates carefully designed visual analytic interventions that facilitate the evaluation of both privacy risks and potential benefits of combining datasets, allowing users to make informed decisions about multi-way joins.

Addressing analytical uncertainty in net load forecasting: Shifting our focus to domain-specific applications of uncertainty mitigation, we developed `Forte`, a visual analytics application for addressing analytical uncertainty in net load forecasting. This system enables energy scientists and grid operators to assess net load variability, analyze the effects of input variables on model performance, and evaluate forecast uncertainties under various conditions. This tool provides a broad understanding of various aspects related to net load forecasting, allowing users to compare model forecasts with actual net load values across different seasons and prediction horizons. It also offers insights into the impact of variables like temperature, humidity, and apparent power on net load forecasts, thus providing a tool for energy planners and grid operators to make more informed decisions. Additionally, `Forte` enables stakeholders to understand how models react to noisy inputs, ensuring that the system is effective in addressing the"known unknowns". This tool demonstrates the power of visual analytics in enhancing the interpretability and reliability of complex AI models in critical domains such as energy planning.

Enhancing trust in AI models for net load forecasting: In our final technical chapter, we extended the work on `Forte` to focus specifically on enhancing trust in AI models for net load forecasting. This extension facilitates the comparison of multiple models across various parameters, including solar penetration levels, dataset resolutions, and different times of day. By enabling users to compare forecasting models with reference models, this work provides a framework for evaluating model performance and building confidence in results. This chapter underscores the importance of trust and transparency in the deployment of AI models in critical infrastructure planning.

The landscape of work presented in this dissertation spans from broad privacy concerns in open data ecosystems to specific applications in energy forecasting. Throughout these chapters, we have demonstrated the effectiveness of visual analytics

in mitigating complex challenges tied to analytical uncertainty. Our work has consistently emphasized the importance of human-in-the-loop approaches, where interactive visualizations provide transparency and enable informed decision-making. Looking to the future, this research opens up several promising avenues for further investigation. There is a need for continued development of privacy-preserving techniques that can adapt to evolving threats and data environments. The integration of machine learning and AI with privacy-preserving visualizations presents both challenges and opportunities for enhancing data protection and utility. Additionally, the application of these techniques to other domains beyond energy forecasting could yield valuable insights and tools for a wide range of data-driven fields. Ultimately, this dissertation contributes to the broader goal of creating interactive workflows that mitigate analytical uncertainty. We envision a data ecosystem that balances openness and transparency with robust privacy protections. By providing tools and frameworks for understanding and mitigating privacy risks, evaluating data utility, and ultimately addressing analytical uncertainty, we aim to empower stakeholders at all levels to make more informed decisions about data sharing, analysis, and application. As our digital world continues to evolve, the principles and approaches developed in this work will serve as important building blocks for future research and practical implementations in privacy-preserving data visualization and analysis.

REFERENCES

[1] A. Dasgupta, R. Kosara, and M. Chen, "Guess Me If You Can: A Visual Uncertainty Model for Transparent Evaluation of Disclosure Risks in Privacy-Preserving Data Visualization," *VizSec*, pp. 1–10, 2019.

[2] N. Heckert, J. Filliben, C. Croarkin, B. Hembree, W. Guthrie, P. Tobias, and J. Prinz, "Handbook 151: NIST/SEMATECH e-Handbook of Statistical Methods," 2002-11-01 00:11:00 2002.

[3] R. A. Pielke Jr., "Who Decides? Forecasts and Responsibilities," *Applied Behavioral Science Review*, vol. 7, no. 2, pp. 83–101, 1999.

[4] P. Rotella, "Is Data The New Oil?" https://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/?sh=55f255b37db3, Apr 2012, (Accessed on 09/05/2021).

[5] A. Ng, "Andrew Ng: Why AI Is the New Electricity — Stanford Graduate School of Business," https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity, 03 2017, (Accessed on 10/08/2024).

[6] J. Trenker, S. S. Menon, N. Tavva, and C. Blumtritt, "If Big Data is the new oil, AI is the new electricity," https://www.statista.com/site/insights-compass-ai-ai-market-overview, 07 2023, (Accessed on 10/08/2024).

[7] "Open Definition 2.1 - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge," https://opendefinition.org/od/2.1/en/, August 2015, (Accessed on 07/19/2021).

[8] D. Linders and S. C. Wilson, "What is open government? one year after the directive," in *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, ser. dg.o '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 262–271.

[9] O. D. Charter, "Our history - International Open Data Charter," https://opendatacharter.net/our-history/, 2020, (Accessed on 07/19/2021).

[10] "NYC Open Data," https://opendata.cityofnewyork.us/, 2021, (Accessed on 10/05/2021).

[11] "Open Data Kansas City," https://data.kcmo.org/, 2021, (Accessed on 10/05/2021).

[12] "City of Dallas Open Data," https://www.dallasopendata.com/, 2021, (Accessed on 10/05/2021).

[13] C. Culnane, B. I. P. Rubinstein, and V. Teague, "Health data in an open world," *arXiv preprint arXiv:1712.05627*, 2017.

[14] A. Lavrenovs and K. Podins, "Privacy violations in Riga open data public transport system," in *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*. IEEE, 2016, pp. 1–6.

[15] J. Yim, R. Chopra, T. Spitz, J. Winkens, A. Obika, C. Kelly, H. Askham, M. Lukic, J. Huemer, K. Fasler *et al.*, "Predicting conversion to wet age-related macular degeneration using deep learning," *Nature Medicine*, vol. 26, no. 6, pp. 892–899, 2020.

[16] N. Woloszko, "Tracking activity in real time with Google Trends," *OECD*, vol. 2020, no. 1634, 2020.

[17] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, "Adaptive Power System Emergency Control using Deep Reinforcement Learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1171–1182, 2019.

[18] DOE Office of Cybersecurity, Energy Security, and Emergency Response, "DOE Delivers Initial Risk Assessment on Artificial Intelligence for Critical Energy Infrastructure — Department of Energy," https://www.energy.gov/ceser/articles/doe-delivers-initial-risk-assessment-artificial-intelligence-critical-energy, 04 2024, (Accessed on 10/08/2024).

[19] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The Role of Uncertainty, Awareness, and Trust in Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 240–249, 2015.

[20] J. Rode, C. Johansson, P. DiGioia, K. Nies, D. H. Nguyen, J. Ren, P. Dourish, D. Redmiles *et al.*, "Seeing Further: Extending Visualization as a Basis for Usable Security," in *Proceedings of the Second Symposium on Usable Privacy and Security*. Pittsburgh, PA, USA: ACM, 2006, pp. 145–155.

[21] J. Montemayor, A. Freeman, J. Gersh, T. Llanso, and D. Patrone, "Information Visualization for Rule-based Resource Access Control," in *Proceedings of the Second Symposium on Usable Privacy and Security*. Pittsburgh, PA, USA: ACM, 2006, pp. 24–25.

[22] A. Dasgupta, J.-Y. Lee, R. Wilson, R. A. Lafrance, N. Cramer, K. Cook, and S. Payne, "Familiarity Vs Trust: A Comparative Study of Domain Scientists' Trust in Visual Analytics and Conventional Analysis Methods," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 271–280, 2016.

[23] A. R. Kandakatla, V. Chandan, S. Kundu, I. Chakraborty, K. Cook, and A. Dasgupta, "Towards trust-augmented visual analytics for data-driven energy modeling," in *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE, 2020, pp. 16–21.

[24] K. Bhattacharjee, M. Chen, and A. Dasgupta, "Privacy-Preserving Data Visualization: Reflections on the State of the Art and Research Opportunities," in *Computer Graphics Forum*, vol. 39. Norrköping, Sweden: Wiley Online Library, 2020, pp. 675–692.

[25] K. Bhattacharjee and A. Dasgupta, "Power to the Data Defenders: Human-Centered Disclosure Risk Calibration of Open Data," in *Proceedings of Symposium on Usable Security and Privacy (USEC) 2023*. San Diego, CA, USA: Symposium on Usable Security and Privacy (USEC) 2023, Feb 2023, pp. 1–6.

[26] K. Bhattacharjee, A. Islam, J. Vaidya, and A. Dasgupta, "PRIVEE: A Visual Analytic Workflow for Proactive Privacy Risk Inspection of Open Data," in *2022 IEEE Symposium on Visualization for Cyber Security (VizSec)*, IEEE. Oklahoma City, USA: IEEE, 2022, pp. 1–11.

[27] K. Bhattacharjee and A. Dasgupta, "VALUE: Visual Analytics driven Linked data Utility Evaluation," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, ser. HILDA '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3597465.3605225

[28] K. Bhattacharjee, S. Kundu, I. Chakraborty, and A. Dasgupta, "Forte: An Interactive Visual Analytic Tool for Trust-Augmented Net Load Forecasting," in *2024 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, IEEE. Washington D.C., USA: IEEE, 2024, pp. 1–5.

[29] ——, "Who should I trust? A Visual Analytics Approach for Comparing Net Load Forecasting Models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21299

[30] National Institute of Standards and Technology, NIST Physics Laboratory, *The NIST Reference on Constants, Units and Uncertainty*. Gaithersburg, MD, USA: National Institute of Standards and Technology, 1998.

[31] Office for National Statistics, "Uncertainty and how we measure it for our surveys - Office for National Statistics," https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/uncertaintyandhowwemeasureit, 01 2020, (Accessed on 10/11/2024).

[32] ——, "Non-financial business economy, UK: Sections A to S," https://www.ons.gov.uk/businessindustryandtrade/business/businessservices/datasets/uknonfinancialbusinesseconomyannualbusinesssurveysectionsas, Office for National Statistics, United Kingdom, 2024, dataset (Accessed on 07/09/2024).

[33] ——, "Non-financial business economy, UK: quality measures," https://tinyurl.com/ms33jdhd, 2024, dataset (Accessed on 07/09/2024).

[34] R. Deonarine, G. Pickering, and Z. Slade, "Assessing and Communicating Uncertainty Toolkit," https://analystsuncertaintytoolkit.github.io/UncertaintyWeb/index.html, 3 2020, (Accessed on 06/14/2024). [Online]. Available: https://analystsuncertaintytoolkit.github.io/UncertaintyWeb/index.html

[35] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

[36] Technology Transformation Services, US General Services Adminstration, "Data.gov home — data.gov," https://data.gov/, 12 2009, (Accessed on 10/11/2024).

[37] Greater London Authority, "London Datastore – Greater London Authority," https://data.london.gov.uk/, 01 2010, (Accessed on 10/11/2024).

[38] "NYC Open Data," https://opendata.cityofnewyork.us/, (Accessed on 10/05/2021).

[39] D. Castellani Ribeiro, H. T. Vo, J. Freire, and C. T. Silva, "An Urban Data Profiler," in *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, 2015, pp. 1389–1394.

[40] "Socrata API," https://dev.socrata.com/, (Accessed on 07/10/2021).

[41] A. Islam and A. Dasgupta, "UrbanForest:Seeing the data forest despite the trees," https://niiv.njitvis.com/assets/publications/2020/urbanForest.pdf, IEEE, 2020, (Accessed on 10/07/2021).

[42] P. Huston, V. Edge, and E. Bernier, "Open science/open data: Reaping the benefits of open data in public health," *Canada Communicable Disease Report*, vol. 45, no. 11, p. 252, 2019.

[43] Centre for Public Impact, "Transportation in New York — Centre For Public Impact (CPI)," https://www.centreforpublicimpact.org/case-study/transportation-in-new-york, 03 2016, (Accessed on 07/14/2024).

[44] S. Bracho, "A Case Study Analysis Assessing Open Data in Urban Passenger Transportation Systems," https://trid.trb.org/View/1393907, 2016, (Accessed on 07/16/2024).

[45] C. Cadwalladr and E. Graham-Harrison, "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach," *The Guardian*, vol. 17, p. 22, 2018.

[46] J. A. Obar and A. Oeldorf-Hirsch, "The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services," *Information, Communication & Society*, vol. 23, no. 1, pp. 128–147, 2020.

[47] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan, "Dark Patterns at Scale," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, p. 1–32, Nov 2019.

[48] R. Motwani and Y. Xu, "Efficient algorithms for masking and finding quasi-identifiers," in *Proceedings of the Conference on Very Large Data Bases (VLDB)*, Vienna, Austria, 2007, pp. 83–93.

[49] L. Sweeney, "Privacy-enhanced linking," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 72–75, 2005.

[50] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, "Publishing data from electronic health records while preserving privacy: A survey of algorithms," *Journal of Biomedical Informatics*, vol. 50, pp. 4–19, 2014.

[51] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[52] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[53] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *21st International Conference on Data Engineering (ICDE'05)*, IEEE. Tokyo, Japan: IEEE, 2005, pp. 217–228.

[54] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, 2006, pp. 24–24.

[55] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, IEEE. Istanbul, Turkey: IEEE, 2007, pp. 106–115.

[56] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[57] H. B. Lee, "Visualization and Differential Privacy," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2017.

[58] K. Nissim, T. Steinke, A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, D. R. O'Brien, and S. Vadhan, "Differential privacy: A primer for a non-technical audience," in *Privacy Law Scholars Conference*, 2017.

[59] Y. Yuan, X. Yuan, H. Wang, M. Tang, and M. Li, "Net load forecasting method in distribution grid planning based on LSTM network," *Science and Technology for Energy Transition*, vol. 79, p. 57, 2024.

[60] X. Wang, H. Wang, B. Bhandari, and L. Cheng, "AI-Empowered Methods for Smart Energy Consumption: A Review of Load Forecasting, Anomaly Detection and Demand Response," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 11, no. 3, pp. 963–993, 2024.

[61] D. Sen, I. Chakraborty, S. Kundu, A. P. Reiman, I. Beil, and A. Eiden, "kPF-AE-LSTM: A Deep Probabilistic Model for Net-Load Forecasting in High Solar Scenarios," *arXiv preprint arXiv:2203.04401*, 2022.

[62] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, "Worst-case background knowledge for privacy-preserving data publishing," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 126–135.

[63] W. Du, Z. Teng, and Z. Zhu, "Privacy-maxent: Integrating Background Knowledge in Privacy Quantification," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. Vancouver, BC, Canada: ACM, 2008, pp. 459–472.

[64] G. T. Duncan and D. Lambert, "Disclosure-Limited Data Dissemination," *Journal of the American Statistical Assn.*, vol. 81, no. 393, pp. pp. 10–18, 1986.

[65] D. Lambert, "Measures of disclosure risk and harm," *Journal of Official Statistics*, vol. 9, pp. 313–331, 1993.

[66] R. Kang, L. Dabbish, N. Fruchter, and S. Kiesler, ""My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security," in *Eleventh Symposium on Usable Privacy and Security ({SOUPS} 2015)*, 2015, pp. 39–52.

[67] J. S. Olson, J. Grudin, and E. Horvitz, "A study of preferences for sharing and privacy," in *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2005, pp. 1985–1988.

[68] A. W. Orlando and A. J. Rosoff, "The New Privacy Crisis: What's Health Got to Do with It?" *The American Journal of Medicine*, vol. 132, no. 2, pp. 127–128, 2019.

[69] M. Büscher, S.-Y. Perng, and M. Liegl, "Privacy, Security, and Liberty: ICT in Crises," in *Censorship, Surveillance, and Privacy: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2019, pp. 199–217.

[70] M. Y. Vardi, "Are we having an ethical crisis in computing?" *Communications of the ACM*, vol. 62, no. 1, p. 7, 2019.

[71] B. C. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (Csur)*, vol. 42, no. 4, pp. 1–53, 2010.

[72] S. Bu, L. V. Lakshmanan, R. T. Ng, and G. Ramesh, "Preservation of patterns and input-output privacy," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 696–705.

[73] G. J. Annas *et al.*, "HIPAA regulations-a new era of medical-record privacy?" *New England Journal of Medicine*, vol. 348, no. 15, pp. 1486–1490, 2003.

[74] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal, "Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, 2020, pp. 1–13.

[75] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, F. Sun *et al.*, "Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts," *The Lancet Global Health*, 2020.

[76] L. Reichert, S. Brack, and B. Scheuermann, "Privacy-Preserving Contact Tracing of COVID-19 Patients," https://eprint.iacr.org/2020/375.pdf, 2020.

[77] T. Li and N. Li, "On the Tradeoff Between Privacy and Utility in Data Publishing," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris, France: ACM, 2009, pp. 517–526.

[78] E. Bertino, D. Lin, and W. Jiang, "A Survey of Quantification of Privacy Preserving Data Mining Algorithms," *Privacy-Preserving Data Mining*, pp. 183–205, 2008.

[79] S. O. Dyke, E. S. Dove, and B. M. Knoppers, "Sharing health-related data: a privacy test?" *Nature Publishing Journal Genomic Medicine*, vol. 1, p. 16024, 2016.

[80] M. Johnson, S. Egelman, and S. M. Bellovin, "Facebook and Privacy: It's Complicated," in *Proceedings of the Eighth Symposium on Usable Privacy and Security*. Washington D.C., USA: ACM, 2012, p. 9.

[81] A. Mazzia, K. LeFevre, and E. Adar, "The PViz comprehension tool for social network privacy settings," in *Proceedings of the Eighth Symposium on Usable Privacy and Security*. Washington D.C., USA: ACM, 2012, p. 13.

[82] S. J. Rizvi and J. R. Haritsa, "Maintaining Data Privacy in Association Rule Mining," in *Proceedings of the 28th International Conference on Very Large Data Bases*. Hong Kong, China: VLDB Endowment, 2002, pp. 682–693.

[83] T. Wang and L. Liu, "Butterfly: Protecting output privacy in stream mining," in *2008 IEEE 24th International Conference on Data Engineering*. IEEE, 2008, pp. 1170–1179.

[84] G. Conti, M. Ahamad, and J. Stasko, "Attacking Information Visualization System Usability Overloading and Deceiving the Human," in *Proceedings of the 2005 Symposium on Usable Privacy and Security*. Pittsburgh, PA, USA: ACM, 2005, pp. 89–100.

[85] E. D. Ragan, H.-C. Kum, G. Ilangovan, and H. Wang, "Balancing Privacy and Information Disclosure in Interactive Record Linkage with Visual Masking," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montral, QC, Canada: ACM, 2018, p. 326.

[86] G. Yee, "Visualization for Privacy Compliance," in *Proceedings of the 3rd International Workshop on Visualization for Computer Security*. Alexandria, VA, USA: ACM, 2006, pp. 117–122.

[87] C.-H. Kao, C.-H. Hsieh, Y.-F. Chu, Y.-T. Kuang, and C.-K. Yang, "Using data visualization technique to detect sensitive information re-identification problem of real open dataset," *Journal of Systems Architecture*, vol. 80, pp. 85–91, 2017.

[88] F. Xiao, M. Lu, Y. Zhao, S. Menasria, D. Meng, S. Xie, J. Li, and C. Li, "An information-aware visualization for privacy-preserving accelerometer data sharing," *Human-centric Computing and Information Sciences*, vol. 8, no. 1, p. 13, 2018.

[89] B. Gao and B. Berendt, "Circles, Posts and Privacy in Egocentric Social Networks: An Exploratory Visualization Approach," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. IEEE, 2013, pp. 792–796.

[90] P. Elagroudy, M. Khamis, F. Mathis, D. Irmscher, A. Bulling, and A. Schmidt, "Can Privacy-Aware Lifelogs Alter Our Memories?" in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. LBW0244.

[91] M. Anwar, P. W. Fong, X.-D. Yang, and H. Hamilton, "Visualizing Privacy Implications of Access Control Policies in Social Network Systems," in *Data Privacy Management and Autonomous Spontaneous Security*. Springer, 2009, pp. 106–120.

[92] M. Bahrini, N. Wenig, M. Meissner, K. Sohr, and R. Malaka, "HappyPermi: Presenting Critical Data Flows in Mobile Application to Raise User Security Awareness," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, pp. 1–6.

[93] J. Becker, M. Heddier, A. Öksüz, and R. Knackstedt, "The Effect of Providing Visualizations in Privacy Policies on Trust in Data Privacy and Security," in *2014 47th Hawaii International Conference on System Sciences*. IEEE, 2014, pp. 3224–3233.

[94] P. S. Dhotre, A. Bihani, S. Khajuria, and H. Olesen, "Take it or Leave it: Effective Visualization of Privacy Policies," in *Cybersecurity and Privacy*. River Publishers, 2017, pp. 39–64.

[95] J. Oksanen, C. Bergman, J. Sainio, and J. Westerholm, "Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data," *Journal of Transport Geography*, vol. 48, pp. 135–144, 2015.

[96] J.-K. Chou, Y. Wang, and K.-L. Ma, "Privacy Preserving Event Sequence Data Visualization using a Sankey Diagram-like Representation," in *SIGGRAPH ASIA 2016 Symposium on Visualization*. ACM, 2016, p. 1.

[97] J.-K. Chou, C. Bryan, and K.-L. Ma, "Privacy Preserving Visualization for Social Network Data with Ontology Information," in *2017 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2017, pp. 11–20.

[98] J.-K. Chou, Y. Wang, and K.-L. Ma, "Privacy Preserving Visualization: A Study on Event Sequence Data," in *Computer Graphics Forum*, vol. 38. Wiley Online Library, 2019, pp. 340–355.

[99] R. Hongde, W. Shuo, and L. Hui, "Differential privacy data Aggregation Optimizing Method and application to data visualization," in *2014 IEEE Workshop on Electronics, Computer and Applications*. IEEE, 2014, pp. 54–58.

[100] A. Dasgupta, M. Chen, and R. Kosara, "Conceptualizing Visual Uncertainty in Parallel Coordinates," *Computer Graphics Forum*, vol. 31, no. 3pt2, pp. 1015–1024, 2012.

[101] A. Dasgupta and R. Kosara, "Adaptive Privacy-Preserving Visualization Using Parallel Coordinates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2241–2248, 2011.

[102] A. Dasgupta, M. Chen, and R. Kosara, "Measuring Privacy and Utility in Privacy-Preserving Visualization," in *Computer Graphics Forum*, vol. 32. Wiley Online Library, 2013, pp. 35–47.

[103] J.-K. Chou and C.-K. Yang, "Obfuscated volume rendering," *The Visual Computer*, vol. 32, no. 12, pp. 1593–1604, 2016.

[104] S. A. Osia, A. S. Shamsabadi, S. Sajadmanesh, A. Taheri, K. Katevas, H. R. Rabiee, N. D. Lane, and H. Haddadi, "A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics," *IEEE Internet of Things Journal*, 2020.

[105] I. Liccardi, A. Abdul-Rahman, and M. Chen, "I Know Where You Live: Inferring Details of People's Lives by Visualizing Publicly Shared Location Data," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose, CA, USA, 2016, pp. 1–12.

[106] N. Andrienko, G. Andrienko, G. Fuchs, and P. Jankowski, "Scalable and Privacy-respectful Interactive Discovery of Place Semantics from Human Mobility Traces," *Information Visualization*, vol. 15, no. 2, pp. 117–153, 2016.

[107] B. Ljubic, D. Gligorijevic, J. Gligorijevic, M. Pavlovski, and Z. Obradovic, "Social network analysis for better understanding of influenza," *Journal of Biomedical Informatics*, vol. 93, p. 103161, 2019.

[108] D. Gotz and D. Borland, "Data-driven Healthcare: Challenges and Opportunities for Interactive Visualization," *IEEE Computer Graphics and Applications*, vol. 36, no. 3, pp. 90–96, 2016.

[109] A. Dasgupta and R. Kosara, "Privacy-Preserving Data Visualization using Parallel Coordinates," in *Visualization and Data Analysis 2011*, vol. 7868. International Society for Optics and Photonics, 2011, p. 78680O.

[110] J. Deeb-Swihart, A. Endert, and A. Bruckman, "Understanding Law Enforcement Strategies and Needs for Combating Human Trafficking," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland, UK: ACM, 2019, p. 331.

[111] X. Wang, T. Gu, X. Luo, X. Cai, T. Lao, W. Chen, Y. Wu, J. Yu, and W. Chen, "A User Study on the Capability of Three Geo-Based Features in Analyzing and Locating Trajectories," *IEEE Transactions on Intelligent Transportation Systems*, 2018.

[112] Y. Takano, S. Ohta, T. Takahashi, R. Ando, and T. Inoue, "Mindyourprivacy: Design and Implementation of a Visualization System for Third-Party Web Tracking," in *2014 Twelfth Annual International Conference on Privacy, Security and Trust*. IEEE, 2014, pp. 48–56.

[113] J. Muchagata, P. Vieira-Marques, and A. Ferreira, "mHealth Applications: Can User-adaptive Visualization and Context Affect the Perception of Security and Privacy?" in *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, Haraklion, Greece, 2019, pp. 444–451.

[114] K. Ghazinour, M. Majedi, and K. Barker, "A Model for Privacy Policy Visualization," in *2009 33rd Annual IEEE International Computer Software and Applications Conference*, vol. 2. IEEE, 2009, pp. 335–340.

[115] X. Wang, J.-K. Chou, W. Chen, H. Guan, W. Chen, T. Lao, and K.-L. Ma, "A Utility-aware Visual Approach for Anonymizing Multi-attribute Tabular Data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 351–360, 2017.

[116] Y. Wang, L. Gou, A. Xu, M. X. Zhou, H. Yang, and H. Badenes, "Veilme: An Interactive Visualization Tool for Privacy Configuration of Using Personality Traits," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul, Republic of Korea: ACM, 2015, pp. 817–826.

[117] X. Wang, W. Chen, J.-K. Chou, C. Bryan, H. Guan, W. Chen, R. Pan, and K.-L. Ma, "Graphprotector: A Visual Interface for Employing and Assessing Multiple Privacy Preserving Graph Algorithms," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 193–203, 2018.

[118] S.-Y. Kung, "Discriminant component analysis for privacy protection and visualization of big data," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 3999–4034, 2017.

[119] C. North, "Toward Measuring Visualization Insight," *IEEE Computer Graphics and Applications*, vol. 26, no. 3, pp. 6–9, 2006.

[120] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky, "Defining Insight for Visual Analytics," *IEEE Computer Graphics and Applications*, vol. 29, no. 2, pp. 14–17, 2009.

[121] W. S. Cleveland and R. McGill, "Graphical perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984.

[122] J. Bertin, *Graphics and Graphic Information Processing*. Walter de Gruyter, 2011, https://doi.org/10.1515/9783110854688.

[123] M. Brehmer and T. Munzner, "A Multi-Level Typology of Abstract Visualization Tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2376–2385, 2013.

[124] D. A. Keim, "Designing pixel-oriented visualization techniques: Theory and applications," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 59–78, 2000.

[125] A. Dasgupta, E. Maguire, A.-R. Alfie, and M. Chen, "Opportunities and Challenges for Privacy-Preserving Visualization of Electronic Health Record Data," in *Proceedings of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records*, Chicago, IL, USA, 2014.

[126] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen, "Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications," in *Eurographics (STARs)*, 2013, pp. 39–63.

[127] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying Personal Genomes by Surname Inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[128] D. George Heilneijer, *The Heilmeijer Catechism*, https://www.darpa.mil/work-with-us/heilmeier-catechism, 1977 (accessed January 23, 2020). [Online]. Available: https://www.darpa.mil/work-with-us/heilmeier-catechism

[129] H.-C. Kum, E. D. Ragan, G. Ilangovan, M. Ramezani, Q. Li, and C. Schmit, "Enhancing Privacy through an Interactive On-demand Incremental Information Disclosure Interface: Applying {Privacy-by-Design} to Record Linkage," in *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019, pp. 175–189.

[130] J.-K. Chou, C. Bryan, J. Li, and K.-L. Ma, "An Empirical Study on Perceptually Masking Privacy in Graph Visualizations," in *2018 IEEE Symposium on Visualization for Cyber Security (VizSec)*. IEEE, 2018, pp. 1–8.

[131] A. Kale, M. Kay, and J. Hullman, "Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019, pp. 1–14.

[132] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," in *Proceedings of the 1996 IEEE Symposium on Visual Languages*. Washington D.C., USA: IEEE, 1996, pp. 336–343.

[133] H. Lin and N. W. Bergmann, "IoT Privacy and Security Challenges for Smart Home Environments," *Information*, vol. 7, no. 3, p. 44, 2016.

[134] M. Gaboardi, J. Honaker, G. King, J. Murtagh, K. Nissim, J. Ullman, and S. Vadhan, "Psi ({\Psi}): a Private data Sharing Interface," *arXiv preprint arXiv:1609.04340*, 2016.

[135] B. P. Hejblum, G. M. Weber, K. P. Liao, N. P. Palmer, S. Churchill, N. A. Shadick, P. Szolovits, S. N. Murphy, I. S. Kohane, and T. Cai, "Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes," *Scientific Data*, vol. 6, no. 1, pp. 1–11, 2019.

[136] C. Anhalt-Depies, J. L. Stenglein, B. Zuckerberg, P. M. Townsend, and A. R. Rissman, "Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science," *Biological Conservation*, vol. 238, no. 108195, 2019.

[137] R. Jia, F. C. Sangogboye, T. Hong, C. Spanos, and M. B. Kjærgaard, "PAD: Protecting Anonymity in Publishing Building Related Datasets," in *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, Delft, Netherlands, 2017, pp. 1–10.

[138] R. Barcellos, J. Viterbo, L. Miranda, F. Bernardini, C. Maciel, and D. Trevisan, "Transparency in practice: using visualization to enhance the interpretability of open data," in *Proceedings of the 18th Annual International Conference on Digital Government Research*, Staten Island, NY, USA, 2017, pp. 139–148.

[139] S. A. Thompson and C. Warzel, *Twelve Million Phones, One Dataset, Zero Privacy*, https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html, 2019 (accessed January 23, 2020).

[140] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing NYC Taxi Data: Does It Matter?" in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2016, pp. 140–148.

[141] G. Andrienko and N. Andrienko, "Privacy Issues in Geospatial Visual Analytics," in *Advances in Location-Based Services: 8th International Symposium on Location-Based Services, Vienna 2011*. Springer, 2012, pp. 239–246.

[142] M. Loughlin and A. Adnane, "Privacy and trust in smart camera sensor networks," in *2015 10th International Conference on Availability, Reliability and Security*. IEEE, 2015, pp. 244–248.

[143] S. Tonyali, K. Akkaya, N. Saputro, A. S. Uluagac, and M. Nojoumian, "Privacy-Preserving Protocols for Secure and Reliable Data Aggregation in IoT-Enabled Smart Metering Systems," *Future Generation Computer Systems*, vol. 78, pp. 547–557, 2018.

[144] S. Dennis, P. Garrett, H. Yim, J. Hamm, A. F. Osth, V. Sreekumar, and B. Stone, "Privacy versus open science," *Behavior Research Methods*, vol. 51, no. 4, pp. 1839–1848, 2019.

[145] M. Correll, "Ethical dimensions of visualization research," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019, pp. 1–13.

[146] R. Y. Wong and D. K. Mulligan, "Bringing Design to the Privacy Table: Broadening "Design" in "Privacy by Design" Through the Lens of HCI," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019, pp. 1–17.

[147] B. Yu and C. T. Silva, "FlowSense: A Natural Language Interface for Visual Data Exploration Within a Dataflow System," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1–11, 2019.

[148] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko, "Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 672–681, 2018.

[149] S. Liu, X. Wang, C. Collins, W. Dou, F. Ouyang, M. El-Assady, L. Jiang, and D. A. Keim, "Bridging Text Visualization and Mining: A Task-Driven Survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 7, pp. 2482–2504, 2018.

[150] N. Newman, "How big data enables economic harm to consumers, especially to low-income and other vulnerable sectors of the population," *Journal of Internet Law*, vol. 18, no. 6, pp. 11–23, 2014.

[151] J. Srinivasan, S. Bailur, E. Schoemaker, and S. Seshagiri, "The Poverty of Privacy: Understanding Privacy Trade-Offs from Identity Infrastructure Users in India," *International Journal of Communication*, vol. 12, pp. 1228–1247, 2018.

[152] J. L. Hicks, T. Althoff, P. Kuhar, B. Bostjancic, A. C. King, J. Leskovec, and S. L. Delp, "Best practices for analyzing large-scale health data from wearables and smartphone apps," *Nature Publishing Journal Digital Medicine*, vol. 2, no. 1, pp. 1–12, 2019.

[153] E. M. Peck, S. E. Ayuso, and O. El-Etr, "Data is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019, pp. 1–12.

[154] Y.-A. De Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[155] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, no. 1, pp. 1–5, 2013.

[156] H. Zang and J. Bolot, "Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study," in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, Rome, Italy, 2011, pp. 145–156.

[157] V. Sekara, L. Alessandretti, E. Mones, and H. Jonsson, "Temporal and cultural limits of privacy in smartphone app usage," *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021.

[158] P. Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review*, vol. 57, p. 1701, 2009.

[159] I. S. Rubinstein and W. Hartzog, "Anonymization and risk," *Washington Law Review*, vol. 91, p. 703, 2016.

[160] M. Zenko, *Red Team: How to Succeed by Thinking Like the Enemy.* Basic Books, 2015.

[161] E. M. Hutchins, M. J. Cloppert, R. M. Amin *et al.*, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains," *Leading Issues in Information Warfare & Security Research*, vol. 1, no. 1, p. 80, 2011.

[162] "Whole Person Care Demographics 2 — SMC Datahub," https://datahub.smcgov.org/dataset/Whole-Person-Care-Demographics-2/qqdq-93h5, (Accessed on 10/07/2021).

[163] "SMC Datahub," https://datahub.smcgov.org/, (Accessed on 10/07/2021).

[164] "Demographics for Public Health, Policy, and Planning — SMC Datahub," https://datahub.smcgov.org/Health-Human-Services/Demographics-for-Public-Health-Policy-and-Planning/kq44-9x3u, (Accessed on 10/07/2021).

[165] "Overdose Information Network Data CY January 2018 - Current Monthly County State Police — PA Open Data Portal," https://data.pa.gov/Opioid-Related/Overdose-Information-Network-Data-CY-January-2018-/hbkk-dwy3, (Accessed on 05/28/2021).

[166] "PA Open Data Portal," https://data.pa.gov/, (Accessed on 11/11/2022).

[167] Y. Dong, K. Takeoka, C. Xiao, and M. Oyamada, "Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, IEEE. Chania, Greece: IEEE, 2021, pp. 456–467.

[168] P. H. Chia, D. Desfontaines, I. M. Perera, D. Simmons-Marengo, C. Li, W.-Y. Day, Q. Wang, and M. Guevara, "KHyperLogLog: Estimating Reidentifiability and Joinability of Large Data at Scale," in *2019 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2019, pp. 350–364.

[169] "City of Fort Lauderdale Police Department Open Data," https://fortlauderdale.data.socrata.com/, (Accessed on 10/05/2021).

[170] "City of New Orleans — Open Data," https://datadriven.nola.gov/home/, (Accessed on 11/02/2021).

[171] "Albany Police — Open Data — City of Albany," https://data.albanyny.gov/, (Accessed on 10/21/2021).

[172] E. F. Codd, "Further Normalization of the Data Base Relational Model," *Data Base Systems*, vol. 6, pp. 33–64, 1972.

[173] "CKAN - The open source data management system," https://ckan.org/, (Accessed on 11/11/2022).

[174] "DKAN Open Data Platform," https://getdkan.org/, (Accessed on 11/11/2022).

[175] B. Ricker, J. Cinnamon, and Y. Dierwechter, "When open data and data activism meet: An analysis of civic participation in Cape Town, South Africa," *The Canadian Geographer/Le Géographe canadien*, vol. 64, no. 3, pp. 359–373, 2020.

[176] S. Baack, "Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism," *Big Data & Society*, vol. 2, no. 2, pp. 1–11, 2015.

[177] S. Milan and M. G. Almazor, "Citizens' media meets big data: The emergence of data activism," *Mediaciones*, vol. 11, no. 14, pp. 120–133, 2015.

[178] A. Bakarov, "A Survey of Word Embeddings Evaluation Methods," *arXiv preprint arXiv:1801.09536*, 2018.

[179] F. Almeida and G. Xexéo, "Word Embeddings: A Survey," *arXiv preprint arXiv:1901.09069*, 2019.

[180] E. AI, "spaCy · Industrial-strength Natural Language Processing in Python," https://spacy.io/, 2015, (Accessed on 11/14/2021).

[181] T. Thongtan and T. Phienthrakul, "Sentiment Classification using Document Embeddings trained with Cosine Similarity," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 407–414.

[182] A. M. Dai, C. Olah, and Q. V. Le, "Document embedding with paragraph vectors," *arXiv preprint arXiv:1507.07998*, 2015.

[183] L. Van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[184] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, LA, USA, 2006, pp. 1027–1035.

[185] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Portland, OR, USA, 1996, pp. 226–231.

[186] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[187] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *ACM Sigmod Record*, vol. 25, no. 2, pp. 103–114, 1996.

[188] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," *ACM Sigmod Record*, vol. 28, no. 2, pp. 49–60, 1999.

[189] E. Schubert and M. Gertz, "Improving the Cluster Structure Extracted from OPTICS Plots," in *Proceedings of the Conference Lernen, Wissen, Daten, Analysen (LWDA)*, Mannheim, Germany, 2018.

[190] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[191] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[192] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 224–227, 1979.

[193] S. G. Eick and A. F. Karr, "Visual Scalability," *Journal of Computational and Graphical Statistics*, vol. 11, no. 1, pp. 22–43, 2002.

[194] T. M. Cover, *Elements of Information Theory*. John Wiley & Sons, 1999.

[195] A. Serjantov and G. Danezis, "Towards an Information Theoretic Metric for Anonymity," in *International Workshop on Privacy Enhancing Technologies*. Springer, 2002, pp. 41–53.

[196] C. Diaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *International Workshop on Privacy Enhancing Technologies*. Springer, 2002, pp. 54–68.

[197] A. Oganian and J. Domingo Ferrer, "A posteriori disclosure risk measure for tabular data based on conditional entropy," *Statistics and Operations Research Transactions*, vol. 27, no. 2, pp. 175–190, 2003.

[198] M. Alfalayleh and L. Brankovic, "Quantifying Privacy: A Novel Entropy-Based Measure of Disclosure Risk," in *International Workshop on Combinatorial Algorithms*, Springer. Duluth, MN, USA: Springer, 2014, pp. 24–36.

[199] T. M. Cover, J. A. Thomas *et al.*, "Entropy, Relative Entropy and Mutual Information," *Elements of Information Theory*, vol. 2, no. 1, pp. 12–13, 1991.

[200] A. Dasgupta, H. Wang, O. Nancy, and S. Burrows, "Separating the Wheat from the Chaff: Comparative Visual Cues for Transparent Diagnostics of Competing Models," *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[201] R. Kosara, F. Bendix, and H. Hauser, "Parallel sets: Interactive exploration and visual analysis of categorical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, pp. 558–568, 2006.

[202] E. Zhu, D. Deng, F. Nargesian, and R. J. Miller, "JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes," in *Proceedings of the 2019 International Conference on Management of Data*. Amsterdam, The Netherlands: ACM, 2019, pp. 847–864.

[203] T. Vos, S. S. Lim, C. Abbafati, K. M. Abbas, M. Abbasi, M. Abbasifard, M. Abbasi-Kangevari, H. Abbastabar, F. Abd-Allah, A. Abdelalim *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[204] D. P. Ballou and H. L. Pazer, "Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 1, pp. 240–243, 2003.

[205] E. Kenneally and K. Claffy, "An internet data sharing framework for balancing privacy and utility," in *Engaging Data: First International Forum on the Application and Management of Personal Electronic Information*. Cambridge, MA, USA: IEEE, 2009, pp. 1–8.

[206] B. Bhumiratana and M. Bishop, "Privacy Aware Data Sharing: Balancing the Usability and Privacy of Datasets," in *Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments*. Corfu, Greece: ACM, 2009, pp. 1–8.

[207] M. Noshad, J. Choi, Y. Sun, A. Hero III, and I. D. Dinov, "A data value metric for quantifying information content and utility," *Journal of Big Data*, vol. 8, no. 1, p. 82, 2021.

[208] Y. Gong, Z. Zhu, S. Galhotra, and R. Castro Fernandez, "Niffler: A Reference Architecture and System Implementation for View Discovery over Pathless Table Collections by Example," *arXiv preprint arXiv:2106.01543*, 2021.

[209] T. Cong, J. Gale, J. Frantz, H. Jagadish, and Ç. Demiralp, "WarpGate: A Semantic Join Discovery System for Cloud Data Warehouses," in *13th Annual Conference on Innovative Data Systems Research (CIDR '23)*. Amsterdam, The Netherlands: CIDR, 2023, pp. 1–7.

[210] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[211] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*. Munich, Germany: Cambridge University Press Cambridge, 2008, vol. 39.

[212] M. Bachmann, "The Levenshtein Python C extension module contains functions for fast computation of Levenshtein distance and string similarity," https://github.com/maxbachmann/Levenshtein, 12 2022, (Accessed on 03/27/2023).

[213] V. Levenshtein, "Binary codes with correction for deletions, insertions and substitutions of characters," *Reports of USSR Academy of Sciences, 163.4: 845*, vol. 163, pp. 845–848, 1965.

[214] H. Hyyrö, Y. J. Pinzón, and A. Shinohara, "New bit-parallel indel-distance algorithm," *Lecture Notes in Computer Science*, vol. 3503, pp. 380–390, 2005.

[215] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[216] R. W. Hamming, "Error Detecting and Error Correcting Codes," *The Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950.

[217] J. Guerra, "Navio: A d3 visualization widget to help summarizing, exploring and navigating large network visualizations," https://github.com/john-guerra/navio, (Accessed on 10/11/2022).

[218] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller, "A Systematic Review on the Practice of Evaluating Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2818–2827, 2013.

[219] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Cambridge, UK: Cambridge University Press, 1993.

[220] T. Koshy, *Catalan Numbers with Applications*. New York, NY, USA: Oxford University Press, 2008.

[221] F. R. Bernhart, "Catalan, Motzkin, and Riordan numbers," *Discrete Mathematics*, vol. 204, no. 1, pp. 73–112, 1999, selected papers in honor of Henry W. Gould.

[222] I. Brito, "A decision model based on expected utility, entropy and variance," *Applied Mathematics and Computation*, vol. 379, pp. 1–21, 2020.

[223] M. Sütçü, "Disutility Entropy in Multi-attribute Utility Analysis," *Computers & Industrial Engineering*, vol. 169, p. 108189, 2022.

[224] T. Asikis and E. Pournaras, "Optimization of privacy-utility trade-offs under informational self-determination," *Future Generation Computer Systems*, vol. 109, pp. 488–499, 2020.

[225] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[226] P. Rane, A. Rao, D. Verma, and A. Mhaisgawali, "Redacting sensitive information from the data," in *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, IEEE. Pune, India: IEEE, 2021, pp. 1–5.

[227] S. B. Basapur, B. Shylaja *et al.*, "Attribute Assailability and Sensitive Attribute Frequency based Data Generalization Algorithm for Privacy Preservation," in *2021 International Conference on Forensics, Analytics, Big Data, Security (FABS)*, vol. 1, IEEE. Bengaluru, India: IEEE, 2021, pp. i–xiv.

[228] R. Ravindra Nikam and R. Shahapurkar, "Data Privacy Preservation and Security Approaches for Sensitive Data in Big Data," in *Recent Trends in Intensive Computing*. Virginia, USA: IOS Press, 2021, pp. 394–408.

[229] New Orleans Police Department, "City of New Orleans — Open Data," https://datadriven.nola.gov/home/, 2019, (Accessed on 11/02/2021).

[230] SETO, "American-Made Net Load Forecasting Prize US DOE," https://www.energy.gov/eere/solar/american-made-net-load-forecasting-prize, 2 2023, (Accessed on 08/01/2023).

[231] J. Dillon, "A Global Look - Residential Solar Adoption Rates," https://www.powermag.com/a-global-look-at-residential-solar-adoption-rates/, 7 2022, (Accessed on 08/01/2023).

[232] P. Lucas and J. Giesen, "Lumen: A Software for the Interactive Visualization of Probabilistic Models Together with Data," *The Journal of Open Source Software*, vol. 63, no. 6, pp. 1–4, 2021.

[233] J. Klaus, M. Blacher, A. Goral, P. Lucas, and J. Giesen, "A visual analytics workflow for probabilistic modeling," *Visual Informatics*, vol. 7, no. 2, 2023.

[234] Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li, and Q. Xia, "Data-Driven Probabilistic Net Load Forecasting with High Penetration of Behind-the-Meter PV," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3255–3264, 2017.

[235] E. Henriksen, U. Halden, M. Kuzlu, and U. Cali, "Electrical Load Forecasting Utilizing an Explainable Artificial Intelligence (XAI) Tool on Norwegian Residential Buildings," in *2022 International Conference on Smart Energy Systems and Technologies (SEST)*. IEEE, 2022.

[236] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[237] J. Tayman and D. A. Swanson, "On the validity of MAPE as a measure of population forecast accuracy," *Population Research and Policy Review*, vol. 18, pp. 299–322, 1999.

[238] M. V. Ellis, "Repeated Measures Designs," *The Counseling Psychologist*, vol. 27, no. 4, pp. 552–578, 1999.

[239] L. M. Sullivan, "Repeated Measures," *Circulation*, vol. 117, no. 9, 2008.

[240] P. G. Sassone and W. A. Schaffer, *Cost-benefit Analysis: A Handbook*. Academic Press New York, 1978, vol. 182.

[241] T. Alaqeel and S. Suryanarayanan, "A comprehensive cost-benefit analysis of the penetration of Smart Grid technologies in the Saudi Arabian electricity infrastructure," *Utilities Policy*, vol. 60, p. 100933, 2019.

[242] D. T. Bui, P. Tsangaratos, V.-T. Nguyen, N. Van Liem, and P. T. Trinh, "Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment," *Catena*, vol. 188, p. 104426, 2020.

[243] S. A. Naghibi and H. R. Pourghasemi, "A Comparative Assessment Between Three Machine Learning Models and Their Performance Comparison by Bivariate and Multivariate Statistical Methods in Groundwater Potential Mapping," *Water Resources Management*, vol. 29, pp. 5217–5236, 2015.

[244] J. Lu, G. Zhou, Q. Fan, D. Zeng, C. Guo, L. Lu, J. Li, C. Xie, C. Lu, F. N. Khan *et al.*, "Performance comparisons between machine learning and analytical models for quality of transmission estimation in wavelength-division-multiplexed systems," *Journal of Optical Communications and Networking*, vol. 13, no. 4, pp. B35–B44, 2021.

[245] A. M. Campbell, S. Kundu, A. Reiman, O. Vasios, I. Beil, and A. Eiden, "Clustering Interval Load with Weather to Create Scenarios of Behind-the-Meter Solar Penetration," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*. Seattle, Washington, USA: IEEE, 2024, pp. 1–5.

[246] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, "The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations," *Computer Graphics Forum*, vol. 39, no. 3, pp. 713–756, 2020.

[247] SETO, "Net Load Forecasting Prize — HeroX," https://www.herox.com/net-load-forecasting, 02 2023, (Accessed on 04/18/2024).

[248] D. Wilks, "Forecast Verification," in *Statistical Methods in the Atmospheric Sciences*, ser. International Geophysics, D. S. Wilks, Ed. NY, USA: Elsevier, 2011, vol. 100, pp. 301–394.

[249] T. DelSole and M. K. Tippett, "Forecast Comparison Based on Random Walks," *Monthly Weather Review*, vol. 144, no. 2, pp. 615–626, 2016.