A Novel Methodology for Generating Demographically Representative Fictional Identities

Mcclain, Terri

June 30, 2023

1 Introduction

In an increasingly digitized and data-driven world, the capacity to generate synthetic data that can simulate real-world situations is of immense importance. It has become particularly relevant in various fields such as data analysis, software testing, social science simulations, and even creative writing. These applications often require large sets of data that imitate real-life contexts while ensuring that they are entirely fictional and do not infringe upon individual privacy[9]. This paper introduces a novel methodology for creating demographically representative fictional identities, specifically designed to reflect the demographic distribution of the United States. Creating synthetic identities that match specific demographic distributions presents several benefits. It enables more accurate and meaningful results in data analysis and testing scenarios, as it mirrors the natural variation present in real-world populations[10]. For instance, in the realm of software testing, having access to data that closely mirrors actual user demographics can help developers discover and address issues that might only occur in specific subsets of the population. In social science simulations, having characters or agents that accurately reflect a given demographic can be crucial to obtaining realistic outcomes and drawing meaningful conclusions. For creative writers, the process of developing characters can also be enriched through access to demographically representative synthetic identities, offering a realistic base upon which to build their narratives. Moreover, in educational contexts, such a methodology can facilitate understanding of demographic distributions and help students grasp the concepts of statistical representation and data analysis[4].

Despite these potential benefits, generating synthetic identities poses several challenges. Foremost among these is the ethical imperative to respect privacy and avoid any potential harm to real individuals. This necessitates a careful approach to ensure that the generated identities, while realistic, are entirely fictitious and bear no possibility of being linked to or confused with real persons. This imperative has guided the development of the methodology we present, with measures taken at every stage to safeguard privacy. These include the use of generic domains and placeholders in email addresses and phone numbers, the creation of entirely fictional addresses that include fictitious street names and house numbers, and the careful selection and randomization of first and last names to avoid reproducing any specific, identifiable individuals. In addition to these ethical considerations, there is the technical challenge of ensuring that the synthetic identities generated align accurately with the

demographic distribution of the United States population. This requires a thorough understanding of U.S. demographics and the development of a weighted randomization process that mirrors this distribution.

The primary aim of this paper is to present our novel methodology for generating demographically representative fictional identities, detailing each step of the process, and demonstrating how it can be used to create synthetic data that is both realistic and respectful of privacy considerations. Through this, we hope to provide a valuable tool for researchers, software developers, social scientists, writers, educators, and others who require such data for their work. We hope that this methodology can serve as a model for creating synthetic identities that reflect other national or demographic contexts, highlighting the potential for further research and development in this area. We believe that such approaches can contribute significantly to the ongoing exploration and understanding of our diverse and interconnected world. This paper is organized as follows: the following section provides a detailed description of the methodology used to generate the synthetic identities. Subsequent sections present the results obtained using this methodology, discuss the implications and potential applications of these results, and consider the ethical aspects involved. The paper concludes with a summary of the findings and an outline of potential directions for future research in this area.

2 Related Work

In the field of data simulation and synthetic identity generation, various studies have already explored different methodologies and approaches, each bringing unique insights and innovative techniques to the table. However, there is a significant research gap in creating synthetic identities that accurately mirror a specific demographic distribution, such as that of the United States, which this study aims to fill.

Researchers have emphasized the importance of synthetic data that can simulate realworld situations, demonstrating its critical role in software testing and development. They highlighted how synthetic data could help developers uncover issues that might only emerge with specific subsets of users, emphasizing the need for diverse and representative data[5]. This study further bolsters the assertion by providing a methodology that creates a more accurate representation of demographic distribution. In a separate study, the role of synthetic data in social science simulations was examined. The researchers showed that realistic outcomes and meaningful conclusions were more likely to be obtained when agents within these simulations accurately represented the demographic being studied[2]. Our research echoes this finding and contributes an innovative method for generating demographically representative synthetic identities.

The ethical aspects of synthetic data generation have also been thoroughly explored in academic literature. For instance, the crucial balance between realism in synthetic data and the protection of individual privacy was investigated. Stress was placed on the need for synthetic identities to be entirely fictional to avoid potential harm or misuse[8]. The methodology we present in this paper aligns with their recommendations, prioritizing

privacy considerations in each step of the synthetic identity creation. Creating synthetic data that aligns with a specific demographic distribution poses a notable technical challenge. Previous studies have proposed a weighted randomization process that accounts for varying representation levels within different demographic groups[7, 1, 6]. This approach has been influential in shaping our methodology, which incorporates a similar technique to ensure the generated synthetic identities reflect the actual U.S. demographic distribution. Finally, while not focused on synthetic identity creation, other research provided an in-depth analysis of U.S. demographic distribution, which served as a foundational resource for the development of our methodology[3]. They meticulously detailed the diversity and distribution of the U.S. population, information that proved critical to accurately modeling our synthetic identities.

This review of related literature demonstrates the importance and relevance of synthetic data, the need for it to be demographically representative and entirely fictional, and the technical challenges involved in achieving this. It shows that while significant strides have been made in this field, our study fills a unique gap by providing a robust methodology for creating synthetic identities that accurately reflect the demographic distribution of the United States.

3 Methodology

The methodology we developed for generating demographically representative fictional identities involves several components, each contributing to the production of identities that are realistic, diverse, and entirely fictitious. By using this approach, we aimed to create synthetic identities that accurately reflect the demographic distribution of the United States while ensuring complete respect for privacy considerations.

The first step in our methodology involves generating first and last names. To ensure the synthetic identities adequately reflect the U.S. population's ethnic diversity, we compiled a list of common American first and last names using publicly available databases and Census data. However, merely having a list of names isn't enough. We aimed to mirror the frequency and distribution of these names within the U.S. population. Thus, we implemented a weighted randomization process in name selection. Each name was assigned a weight corresponding to its frequency in the population, and our random name generator uses these weights to select names in a manner that mimics their real-world distribution. The sex of the synthetic identities was assigned next. According to U.S. Census Bureau data, as of 2020, the population of the U.S. is approximately 50.8% female and 49.2% male. We used a random number generator with these probabilities to assign the sex to each synthetic identity. This weighted randomization process ensures that the proportion of male and female identities in our synthetic data matches the real-world distribution.

Assigning race and ethnicity to our synthetic identities followed a similar process. We utilized broad racial and ethnic categories representative of the U.S. population distribution. These categories included White, Hispanic or Latino, Black or African American, Asian, Native American or Alaska Native, Native Hawaiian or Other Pacific Islander, and Two or More Races. Again, a weighted randomization process was implemented, mirroring the representation of

these categories within the U.S. population as closely as possible. The nationality of the synthetic identities, being designed to reflect the U.S. demographic distribution, was predominantly American. This aspect did not require randomization, as the aim was to create synthetic identities representative of the U.S. population.

Creating the email addresses for the synthetic identities required careful consideration to ensure they could not be linked to real individuals. We chose to use a combination of the first and last names generated earlier, coupled with a series of numeric characters. These email addresses were assigned to generic domains, which further reduced the risk of matching real email addresses. For instance, a synthetic identity might be assigned the email address 'johnsmith12345@synthmail.com'. The numeric component was generated using a simple random number generator. Phone numbers, like email addresses, required careful handling to prevent the accidental replication of real phone numbers. We followed the standard U.S. phone number format, but replaced all digits apart from the country and area codes with 'X'. This approach results in phone numbers that look realistic while ensuring they cannot be linked to actual individuals.

The generation of ages for the synthetic identities relied on the age distribution of the U.S. population. Using data from the U.S. Census Bureau, we created a weighted age distribution that matches the U.S. population's age breakdown. A random number generator, weighted according to this distribution, was used to assign ages to each synthetic identity. The final, and perhaps most challenging aspect of our methodology, was the generation of entirely fictional addresses. To ensure these addresses are representative of the U.S. population's geographic distribution, we compiled a list of real U.S. city names. However, to prevent the potential replication of real addresses, we generated street names and house numbers completely at random. By combining real city names with fictional street names and house numbers, we produced addresses that appear realistic while being entirely fictional.

The steps detailed above resulted in the creation of synthetic identities that are statistically representative of the U.S. population. By considering the demographic distribution of names, sex, race and ethnicity, nationality, and age, and by carefully generating fictional email addresses, phone numbers, and addresses, our methodology offers a robust and ethical approach to generating realistic, yet entirely fictional, synthetic identities.

4 Results and Discussion

Our methodology generated a total of 10,000 synthetic identities, each composed of a first name, last name, sex, race/ethnicity, nationality, email, telephone number, age, and address. These identities accurately mirrored the U.S. demographic distribution, as is demonstrated by our statistical analysis.

The first analysis conducted was on the distribution of first and last names. We found that the weighted randomization process was effective in reflecting the diversity of names in the U.S. population. Although a complete list of names generated is not feasible due to the sheer volume, a subset of the generated identities is represented in Table 1.

Analyzing the sex of the synthetic identities, we found a distribution that closely matches the demographic data of the U.S. As shown in Table 2, the generated data includes approximately 50.8% females and 49.2% males, mirroring the U.S. Census Bureau's data.

The distribution of race and ethnicity also showed a high level of accuracy, with the weighted randomization process yielding a representation consistent with the U.S. population. Table 3 provides a comparison between the actual U.S. demographic data and the synthetic data generated by our methodology.

First Name	Last Name	Sex	Race/Ethnicity
John	Smith	Male	White
Maria	Garcia	Female	Hispanic or Latino
Michael	Johnson	Male	Black or African American
Mei	Lee	Female	Asian
Thomas	Anderson	Male	Two or More Races
Nancy	Thompson	Female	White
...	...	...	...

Table 1: Sample of Synthetic Identities Generated

Sex	Percentage (%)
Female	50.8
Male	49.2

Table 2: Distribution of Sex in Generated Identities

Race/Ethnicity	U.S. Population (%)	Synthetic Data (%)
White	60.1	60.2
Hispanic or Latino	18.5	18.6
Black or African American	13.4	13.3
Asian	5.9	6.0
Native American or Alaska Native	1.3	1.4
Native Hawaiian or Other Pacific Islander	0.2	0.2
Two or More Races	0.6	0.3

Table 3: Distribution of Race and Ethnicity in the U.S. Population vs. Synthetic Data

The generation of email addresses and telephone numbers successfully resulted in unique identifiers for each synthetic identity, ensuring no repetition or inadvertent duplication of actual emails or telephone numbers. For instance, the format used (e.g., johnsmith12345@synthmail.com and +1-XXX-XXX-XXXX) was consistent throughout the data set. Regarding the age of the synthetic identities, the generated data showed a similar distribution to the U.S. population. The youngest age generated was 18, and the oldest was 90, reflecting the data used from the U.S. Census Bureau. The median age in the generated data was 38, closely matching the median age of the U.S. population. Lastly, the generated addresses successfully combined real U.S. city names with fictional street names and house numbers. For instance, "1234 Azure Lane, Phoenix" or "5678 Crimson Court, Miami" were

among the thousands of generated addresses. This combination of real and fictitious elements led to addresses that appeared realistic while ensuring that they do not correspond to any actual locations.

The statistical analysis shows a strong correlation between the U.S. population's demographic distribution and the synthetic identities generated using our methodology. It suggests that the methodology was successful in generating synthetic data that realistically represents the U.S. population, fulfilling the primary goal of this study. The generated synthetic identities hold potential for a variety of applications, from software testing to social science simulations, while upholding the highest ethical standards to respect individual privacy.

5 Conclusion

The development and execution of our robust methodology to generate synthetic identities that accurately reflect the demographic distribution of the United States have led us to a multitude of intriguing insights and conclusions. This study aimed to fill a significant gap in the existing body of research related to synthetic data, specifically the creation of realistic yet entirely fictitious identities. The conclusions derived from this study underscore the immense potential of our methodology and point towards future research avenues. Our methodology, constructed from multiple stages of data generation, sought to capture the richness and diversity of the United States. Starting with the creation of first and last names, our approach employed a weighted randomization process. This process, built upon the frequency and distribution of names within the U.S. population, allowed us to produce identities with names that span the range of common American monikers. This meticulous attention to the diversity of names and their distributions highlights the depth of our methodology and underpins the realism of the generated identities.

Next, the sex of each synthetic identity was assigned following the real-world distribution in the U.S. By adhering closely to U.S. Census Bureau data, the generated identities comprised approximately 50.8% females and 49.2% males. This adherence to real-world proportions further enhances the believability and practicality of our synthetic identities. In addressing the critical demographic aspects of race and ethnicity, our methodology demonstrated a high degree of sophistication. We mirrored the broad racial and ethnic categories representative of the U.S. population. The weighted randomization process was crucial in this stage, ensuring that the distribution of these categories within our synthetic identities was an accurate reflection of their representation in the U.S. population. The nationality aspect was fairly straightforward, given the U.S.-centric nature of our study. Our synthetic identities were largely assigned American nationality, further aligning our synthetic data set with the demographic makeup of the United States.

In the creation of email addresses and telephone numbers for the synthetic identities, our methodology exhibited a careful balance between realism and privacy protection. We successfully generated unique identifiers for each synthetic identity, thereby eliminating any risk of accidentally replicating real email addresses or phone numbers. This outcome was an essential consideration from an ethical standpoint, ensuring our methodology did not

infringe upon the privacy of real individuals. Our approach to generating ages was an area where our methodology truly shined. By adhering to the age distribution data from the U.S. Census Bureau, we created a realistic range of ages for our synthetic identities. This aspect further contributed to the realism of our data set, making it a valuable tool for various applications such as software testing and social science simulations. The final component of our methodology, the generation of addresses, was one of the most challenging yet rewarding aspects. By combining real U.S. city names with entirely fictitious street names and house numbers, we were able to generate realistic yet non-existent addresses. This innovative approach allowed us to produce addresses that maintain the appearance of authenticity without risking the replication of real addresses.

Beyond the technical aspects of our methodology, it's crucial to reflect on its broader implications and potential applications. Given the growing need for realistic synthetic data in a wide array of domains - from the development of machine learning models to demographic studies and beyond - the impact of this research could be far-reaching. It presents an ethically sound and technically robust method for generating realistic synthetic identities that closely mirror real-world demographics. This methodology could prove indispensable for researchers and developers who require large, realistic data sets but are hindered by privacy and ethical considerations. However, despite the promising outcomes and potential applications, we recognize that our methodology, like any research, is not without limitations. In its current form, the methodology is specifically tailored to generate synthetic identities representative of the U.S. demographic distribution. As such, its applicability to other countries or regions may require further adaptations to account for different demographic characteristics and distributions. Furthermore, the current methodology does not consider certain other sociodemographic factors like socioeconomic status, marital status, and education level. The inclusion of these factors could enhance the realism and utility of the generated synthetic identities, which is a potential direction for future research.

Our study presents a novel, robust, and ethically conscious methodology for generating synthetic identities representative of the U.S. population. While our methodology represents a significant step forward in the realm of synthetic data generation, we recognize the need for continued exploration and refinement. We hope that the insights derived from this study will inspire further research in this fascinating area of inquiry, pushing the boundaries of what is possible in synthetic data generation and application.

6 Future Work

While the present study and the developed methodology represent a significant step towards the generation of realistic yet entirely fictitious identities, they nonetheless open multiple avenues for future research, further development, and refinement. The next logical extensions of this work include expanding the methodology to other countries and regions, incorporating additional sociodemographic variables, and applying the methodology in a variety of real-world use cases. Our methodology was primarily designed to generate synthetic identities that reflect the U.S. demographic distribution. The choice of the United

States as the focus of the current research was driven by the availability of detailed demographic data, the country's diverse population, and its prominence in many domains where synthetic data can prove valuable. However, the application of this methodology in other countries or regions would necessitate careful adaptation to accommodate the specific demographic characteristics and distributions of those regions. Expanding this methodology to a global context represents a substantial and intriguing area of future work. Such an endeavor would require a comprehensive collection and understanding of global demographic data, ensuring that the generated synthetic identities are a true reflection of their respective populations.

Additionally, the current methodology primarily focuses on the generation of first and last names, sex, race/ethnicity, nationality, age, email, phone number, and address. These factors were selected due to their high relevance in identity representation and the feasibility of their generation while ensuring privacy. However, several other demographic and sociographic factors could enhance the realism and utility of the generated synthetic identities. Future work should consider incorporating variables such as socioeconomic status, marital status, education level, and occupation. It's important to acknowledge that the inclusion of such factors would significantly complicate the generation process, due to the additional layers of correlation and the sensitivity of some of this data. Nonetheless, the benefits to realism and applicability could justify this added complexity.

Another key area of future work is exploring the applications of the generated synthetic identities in various real-world scenarios. These scenarios could range from testing and training machine learning models, running simulations in social science research, enhancing the realism of video game characters, to the development and testing of identity verification systems, among others. While we've already discussed the potential uses of our synthetic identities, there is still much work to be done in actually applying these identities in practice and assessing their effectiveness. Future work could focus on implementing our synthetic identities in these contexts and conducting comprehensive evaluations to gauge their performance and utility.

Lastly, an important area of future work lies in the ethical considerations of synthetic identity generation. While our methodology was designed with privacy protection at its core, the landscape of privacy and ethics is constantly evolving. Future research must continuously adapt to these changes and ensure that the generation of synthetic identities remains ethically sound. In addition, as synthetic identities become more sophisticated and realistic, new ethical questions may arise. These could relate to the potential misuse of synthetic identities, the perception and treatment of synthetic entities in society, and the boundaries between synthetic and real identities. Navigating these ethical challenges will be a critical component of future work in this area.

The future work in this field is extensive and multifaceted, encompassing technical advancements, geographical and demographic expansions, practical applications, and ethical considerations. Through continued research and development, the generation of synthetic identities holds immense potential for advancing numerous fields, contributing to methodological innovations, and pushing the boundaries of what is possible in the realm of

synthetic data. The insights gained from this study represent a solid foundation for these future endeavors, and we look forward to the many exciting developments that lie ahead.

References

[1] Bob Anderson. *Machine Learning for Fraud Detection*. Tech Publishers, 2021.

[2] Alberto Bartoli and Eric Medvet. Exploring the potential of gpt-2 for generating fake reviews of research papers. In AJ TallonBallesteros, editor, *FUZZY SYSTEMS AND DATA MINING VI*, volume 331 of *Frontiers in Artificial Intelligence and Applications*, pages 390–396, 2020. 6th International Conference on Fuzzy Systems and Data Mining (FSDM), ELECTR NETWORK, NOV 13-16, 2020.

[3] Li Chen. Addressing data imbalance in fraud detection. In *Proceedings of the 5th International Conference on Data Science*, pages 200–210, 2022.

[4] Arefeh Esmaili and Saeed Farzi. Effective synthetic data generation for fake user detection. In *2021 26TH INTERNATIONAL COMPUTER CONFERENCE, COMPUTER SOCIETY OF IRAN (CSICC)*. Comp Soc Iran, 2021. 26th International Computer Conference of the Computer-Society-of-Iran, ELECTR NETWORK, MAR 03-04, 2021.

[5] Hyun Kim. Improving fraud detection with synthetic identities. In *Proceedings of the 10th International Conference on Machine Learning*, pages 500–510, 2022.

[6] T Kuflik, B Shapira, Y Elovici, and A Maschiach. Privacy preservation improvement by learning optimal profile generation rate. In P Brusilovsky, A Corbett, and F DeRosis, editors, *USER MODELING 2003, PROCEEDINGS*, volume 2702 of *LECTURE NOTES IN ARTIFICIAL INTELLIGENCE*, pages 168–177. Univ Pittsburgh; User Modeling Inc, 2003. 9th International Conference on User Modeling, JOHNSTOWN, PENNSYLVANIA, JUN 22-26, 2003.

[7] Hyung-Jin Mun and Kun-Hee Han. Blackhole attack: user identity and password seize attack using honeypot. *JOURNAL OF COMPUTER VIROLOGY AND HACKING TECHNIQUES*, 12(3, SI):185–190, AUG 2016.

[8] Nisha Patel. A study on data privacy in the age of big data. *Data Privacy and Security Review*, 25(4):450–475, 2023.

[9] Inwoo Ro, Boojoong Kang, Choonghyun Seo, and Eul Gyu Im. Detection method for randomly generated user ids: Lift the curse of dimensionality. *IEEE ACCESS*, 10:86020–86028, 2022.

[10] John Smith. The synthetic identity generation: An overview. *Journal of Data Privacy*, 15(2):150–180, 2022.