DeepPhish: Effects of Generative Artifacts within Fabricated Profiles on User Trust

OLIVIA FIGUEIRA, Santa Clara University
GANG WANG, University of Illinois at Urbana-Champaign

ABSTRACT
Fake online social network (OSN) profiles created using natural language and image generation models are becoming increasingly sophisticated and thus harder to detect by traditional methods. Current research focuses on improving the detection of intra-field inaccuracies and artifacts, meaning those that exist within a deep fake image or within generated text in isolation [5, 6, 8]. However, personas and OSN profiles are built on a collection of multi-modal fields that must maintain both intra-field and inter-field consistency. This project aims to determine how intra-field and inter-field inconsistencies affect perceived trust by OSN users and how those inconsistencies compare with each other leveraging a user study with fake LinkedIn profiles created using generative text and image algorithms. At the time this report was written, we had not begun the user study, so the results are forthcoming.

1 INTRODUCTION

A considerable share of modern society participates in online platforms that depend on user-provided content. Many of these sites are populated by fabricated content and personas that aim to deceive and exploit the trust of real users to disseminate fake news [15], promote social discord [4], and perform social engineering attacks [9]. Traditional methods for fake profile generation have either required extensive manual effort, which is costly, or made use of templates and stock photos, which are easily detectable, overall limiting the impact of such attacks.

However, developments in generative image and text algorithms may enable malicious users to create more convincing fraudulent profiles without as much manual effort. Through architectural breakthroughs in generative text and image algorithms, such as GPT-2 [13] and StyleGAN2 [7], these generative models can now create images and text with unprecedented levels of quality and realism. This gives rise to concern over an attacker's ability to create realistic and non-templated profiles that are much harder to detect by way of traditional detection methods.

Examples of such profiles already exist on OSNs, such as LinkedIn. According to a report by the Associated Press, there is evidence of a "would-be spy using an AI-generated profile picture to fool contacts on LinkedIn" [16]. Experts in the field of artificially generated images reviewed the profile and have identified all the indicators that it is a generated image, or a deep fake, such as those created by StyleGAN2 [7, 14]. This fake profile is connected to high-profile United States politicians, which is clearly a security risk since experts believe accounts such as this one are run by foreign agents as part of state-run operations [14].

Prior work regarding automated detection focuses on generated text and image in isolation [5, 6, 8] and consequently do not account for semantic inconsistencies across multi-modal fields. Real LinkedIn profiles consist of a collection of fields, such as name, image, textual summary, education, experience, and more, all of which maintain semantic relationships between one another. These fields may be leveraged to distinguish between real and fabricated profiles given that current image and text generation and detection algorithms perform in isolation and do not account for relationships between these fields.

We hypothesize that temporal relationships, such as apparent age in photo and years of experience, professional relationships, such as field of degree and field of job, often exist between fields within real profiles, and inconsistencies within such relations may be indicative of a fabricated profile. We aim to analyze the effects of these inter-field inconsistencies on user trust in addition to traditional intra-field inconsistencies, such as image and text artifacts generated by the respective generative algorithms. We investigate the following research questions in this work:

RQ. 1 Can users detect inter-field inconsistencies when distinguishing between real and fabricated profiles?
RQ. 2 How do inter-field inconsistencies compare with traditional intra-field inconsistencies in relation to user trust and behavior?

2 RELATED WORK

In [10], Ma et al. conduct a user study using Airbnb host profiles to examine user trust in host profiles created using generative text algorithms. They focus on the text-based description included on Airbnb profiles, and not the other fields, such as photos, reviews, and social media verification status, in their user study. In this work, we use several fields of LinkedIn profiles, including photo, name, textual summary, education, and experience sections, to investigate both intra- and inter-field inconsistencies as they relate to user trust and profile trustworthiness.

In [3], Everett et al. conduct a user study to investigate if and how well users can detect automated text online. They find that "typical Internet users are twice as likely to be deceived by automated content than security researchers." Similarly, they only focus on the users' ability to detect automated text in isolation.

In [8], Kennedy et al. propose a method for automated text detection in the context of online reviews. Their work focuses on detecting automated content on text in isolation. Hsu et al. [6] and Guera et al. [5] propose deep fake image and video detection algorithms, respectively, leveraging machine learning methods. These works focus on the detection of fake images in isolation, while in this work we investigate whether users can identify fake profiles that contain deep fakes as one of the inconsistencies and how it compares with the detection of other field inconsistencies.

In [1], Adikari et al. propose a data mining approach to detect fake profiles on LinkedIn using static features of profiles. They verify the results manually by checking for semantic consistency between

Authors' addresses: Olivia Figueira, ofigueira@scu.edu, Santa Clara University; Gang Wang, gangw@illinois.edu, University of Illinois at Urbana-Champaign.

attributes among other features, reuse of information across different accounts, credibility of connections, and more, while we aim to investigate whether users are able to identify these inconsistencies within one single profile.

3 METHOD

To reiterate, the aim of this work is to investigate the following research questions:

RQ. 1 Can users detect inter-field inconsistencies when distinguishing between real and fabricated profiles?

RQ. 2 How do inter-field inconsistencies compare with traditional intra-field inconsistencies in relation to user trust and behavior?

We intend to answer these questions leveraging a user study in which participants are presented a series of LinkedIn profiles created using generative text and image models GPT-2 [13] and StyleGAN2 [7], respectively. Participants will be asked to answer questions related to metrics of trust described in [10], those being ability, benevolence, and integrity. We hypothesize that higher levels of trust indicate greater perceived authenticity in the profile and lower levels of trust indicate a perception of fraudulence [10]. Since we have not yet begun the studies, we will not disclose the study design in greater detail in this report.

3.1 Profile Data

The data used to create the content of the LinkedIn profiles was obtained from MightyRecruiter's résumé database [11]. We chose five occupation groups for which the distributions of age, gender, and ethnicity were well varied using demographic statistics from the United States Department of Labor [12]; computer and information technology, healthcare, sales, arts and design, and legal occupations. We obtained 29,000 publicly available résumés across these five occupations to serve as the base data set for the LinkedIn profiles. The résumés contain many of the same contents as a typical LinkedIn profile, those being name, self summary, job experience, education, and skills. The real names are substituted for names selected based on data from the United States Census on most common names by race and gender [2] in order to remove personally identifiable information from the profiles.

The images in each LinkedIn profile were generated using StyleGAN2 [7]. We selected images that appear to match the given occupation's distribution of race, gender, and age.

3.2 Base Profile Creation

The LinkedIn profiles to be presented to the survey participants contain the following fields: name, photo, current title, summary, experience section, and education section.

In order to create the profiles, we first selected base profiles from each occupation based on the résumé summary. The final base profile summaries must meet these requirements after manual editing:

(1) Reference at least three out of four of these items: current job, education (i.e., degree(s)), skills related to occupation, and years of experience
(2) Be 80-100 words in length
(3) Contain no major grammar issues

We maintain that each summary should contain those four references (i.e., current job, education, skills, and years of experience) because these are the pieces of information that can be cross-referenced with the other elements of the LinkedIn profile. If an original summary only referenced three out of the four references, we added the remaining one using the same sentence structure and content found in other base profile summaries with the given profile's information. If an original summary was too long but contained the necessary references, then we removed the superfluous information. Finally, we fixed any grammar issues to ensure the final base summary met the above requirements.

In order to anonymize the profiles, for each base résumé we replaced the education and experience sections with those of other résumés with similar occupations and educational backgrounds. In addition, we replaced the names of all companies and schools referenced in the profiles with fictional ones in order to increase anonymity and remove any bias that could be associated with these institutions. We followed these criteria in selecting profiles from which to take education and experience sections for the base profiles:

(1) The source education section must contain a degree and field of study that is the same or as similar as possible to that of the base profile.
(2) The source experience section must contain the same or similar occupations involving the same skills mentioned in the base summary, and it should only contain three experience items so the profile is not too long for participants to read during the survey.

Once these changes were made, we updated the referenced years of experience, current job title, and educational experience in the summary to ensure consistency across the profile sections.

Finally, we selected an image from the StyleGAN2 [7] output that does not contain any artifacts and a name, both of which corresponding to the given profiles most prevalent race, gender, and age using data from the U.S. Census [2].

3.3 Profile Variant Creation

Following with our research questions, we created variations of each base profile that contain intra- and inter-field inconsistencies. Each variant profile only varies by one difference so that we may better analyze the survey data and answer our research questions regarding the comparison between intra- and inter-field inconsistencies and whether users can identify inter-field inconsistencies. For each base profile, there are five variations. The profile variations and creation methodology are as follows:

Base Profile

- The profile contains no inconsistencies. Refer back to section 3.2 for base profile creation methodology.

Profiles with Intra-Field Inconsistencies

- The profile image contains artifacts commonly found in deep fakes (i.e. smudges on skin, halo effect around hair, indistinct background, blurred earrings, etc.). This is done by selecting a StyleGAN2 [7] output image that matches the race, gender, and age of the base profile image that contains obvious artifacts.

- The summary text contains artifacts commonly found in generated text (i.e. repetition, grammatical errors, off-topic content, etc.). This is done by running the GPT-2 [13] model trained for the specific given occupation with the first few words of the original summary as a prompt until we generate a summary that contains at least three of the four desired references to education, current job, skills, and years of experience. This way, the summary will contain the same categories of information so that the format of all summaries is consistent, but it will contain artifacts from the generative algorithm. We also made sure the chosen summary only differs from the original summary by ±10 words. We manually selected the summary that reflects these requirements.

Profiles with Inter-Field Inconsistencies

- **Temporal Inconsistency**
  - The years of experience in the summary does not match dates in the experience and education section. We accomplished this by inputting a portion of the summary preceding the reference to the years of experience into the GPT-2 [13] model trained for the specific given occupation and selecting a generated output that contains a number different from the original. The rest of the summary remains the same as the original.
  - The person in profile image appears significantly younger than the implied age based on their indicated years of experience in the summary section. We did this by selecting an image output by the StyleGAN2 [7] algorithm that matches the target gender and race of the original profile but appears to be much younger than the original image. This way, we created an inter-field temporal inconsistency that would require the profile viewer to see that the profile persona appears too young to have the given years of experience listed in the profile, for example.
- **Professional Inconsistency**
  - The summary lists a different field of experience and education than do the education and experience sections of the profile. We created this summary by running the GPT-2 [13] model trained for the specific given occupation several times starting with the first few words of the original summary as prompt. In order to maintain the original summary's grammar and sentence structure while only changing the professional references, we only generated text where those references occur in the original summary, with the original text as prompt. This way, the new summary contains references to a different occupation from what is shown in the rest of the LinkedIn profile, creating an inter-field professional inconsistency.

3.4 Study Design

Since we have not yet begun the user study, we will not disclose the study design in this report.

4 DISCUSSION

As was mentioned before, at the time this report was written we had not begun the user studies and thus the results of the user studies and respective analyses are forthcoming. To reiterate, we hope to answer these research questions:

RQ. 1 Can users detect inter-field inconsistencies when distinguishing between real and fabricated profiles?

RQ. 2 How do inter-field inconsistencies compare with traditional intra-field inconsistencies in relation to user trust and behavior?

Regarding RQ. 1, if users cannot detect inter-field inconsistencies when distinguishing between real and fabricated profiles, this may be a strong reason for the development of consistency detection algorithms to warn OSN users and help OSNs detect suspicious profiles. On the other hand, if users can detect inter-field inconsistencies, attackers would have to consider such relationships when crafting profiles. Regarding RQ. 2, if users can detect both inter- and intra-field inconsistencies, we want to investigate how they respond to each type of profile. If they can detect both but respond to either with varying levels of trust, this would suggest that certain inconsistencies are more important with respect to user trust, and would similarly give both OSNs and attackers more insight into what they should be focusing on for both fraudulent profile detection and creation, respectively.

5 FUTURE WORK

A future project we are interested in pursuing is an analysis of how stereotypical relationships regarding gender and race within fabricated profiles impact user trust and behavior on OSNs. In this work, we used the most demographically prevalent gender and race for each occupation in the United States to select the image and name for each profile. The names were selected using the most common names for a given gender and race combination. We hope to analyze these variables in a very similar context to investigate whether stereotypical relationships and related inconsistencies affect user trust and behavior within OSNs. We want to investigate how user bias may play a role in their trust and behavior of OSN profiles as it relates to the intersection of gender and race with occupation, experience, and education. This future work would be an extension of this project because stereotypical relationships are another field within OSN profiles that may be leveraged in distinguishing between authentic and fabricated profiles.

6 ACKNOWLEDGMENTS

This work is supported by the Computing Research Association's Distributed Research Experience for Undergraduates program, Ph.D. students Jaron Mink and Qingying Hao in Gang Wang's i-DASH Lab at the University of Illinois-Urbana Champaign, and other faculty and student researchers in the Computer Science Department at the University of Illinois-Urbana Champaign.

REFERENCES

[1] S. Adikari and K. Dutta. 2014. Identifying Fake Profiles in LinkedIn. *ArXiv* abs/2006.01381 (2014).

[2] United States Census Bureau. 2016. Frequently Occurring Surnames from the 2010 Census. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

[3] Richard M. Everett, Jason R. C. Nurse, and Arnau Erola. 2016. The Anatomy of Online Deception: What Makes Automated Text Convincing?. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing* (Pisa, Italy) *(SAC*

'16). Association for Computing Machinery, New York, NY, USA, 1115–1120. https://doi.org/10.1145/2851613.2851813

[4] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (Jun 2016), 96–104. https://doi.org/10.1145/2818717

[5] Zhang Hongmeng, Zhu Zhiqiang, Sun Lei, Mao Xiuqing, and Wang Yuehan. 2020. A Detection Method for DeepFake Hard Compressed Videos Based on Super-Resolution Reconstruction Using CNN. In *Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence* (Qingdao, China) *(HPCCT & BDAI 2020)*. Association for Computing Machinery, New York, NY, USA, 98–103. https://doi.org/10.1145/3409501.3409542

[6] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee. 2020. Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences* 10 (01 2020), 370. https://doi.org/10.3390/app10010370

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019. Analyzing and Improving the Image Quality of StyleGAN. arXiv:1912.04958 [cs.CV]

[8] Stefan Kennedy, Niall Walsh, Kirils Sloka, Andrew McCarren, and Jennifer Foster. 2019. Fact or Factitious? Contextualized Opinion Spam Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Florence, Italy, 344–350. https://doi.org/10.18653/v1/P19-2048

[9] Katharina Krombholz, Dieter Merkl, and Edgar Weippl. 2012. Fake identities in social media: A case study on the sustainability of the Facebook business model. *Journal of Service Science Research* 4 (12 2012). https://doi.org/10.1007/s12927-012-0008-z

[10] Xiao Ma, Jeffery T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2397–2409. https://doi.org/10.1145/2998181.2998269

[11] MightyRecruiter. 2020. MightyRecruiter. https://www.mightyrecruiter.com/

[12] U.S. Bureau of Labor Statistics. 2020. Labor Force Statistics from the Current Population Survey. https://www.bls.gov/cps/demographics.htm

[13] A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[14] Raphael Satter. 2019. Experts: Spy used AI-generated face to connect with targets. https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d

[15] Chengcheng Shao, Giovanni Ciampaglia, Onur Varol, Alessandro Flammini, Filippo Menczer, and Kai-Cheng Yang. 2018. The spread of low-credibility content by social bots. *Nature Communications* 9 (11 2018). https://doi.org/10.1038/s41467-018-06930-7

[16] James Vincent. 2019. A spy reportedly used an AI-generated profile picture to connect with sources on LinkedIn. https://www.theverge.com/2019/6/13/18677341/ai-generated-fake-faces-spy-linked-in-contacts-associated-press