Generative Adversarial Networks (GANs)

What are they, why should infosec care, and what to do.

Matt Richard, IRL
April 27, 2022

Whoami

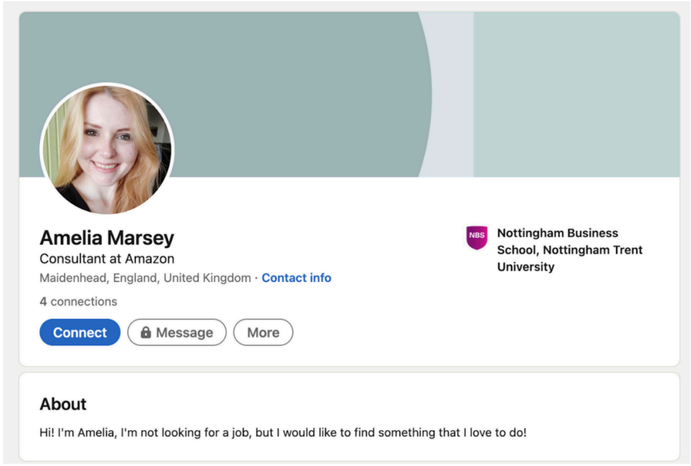
- Raytheon, MITRE, Facebook, Microsoft, IRL
- Reverser, builder, researcher, “threat scientist”
- Fraud, disinfo, APT, malware,

What we will cover

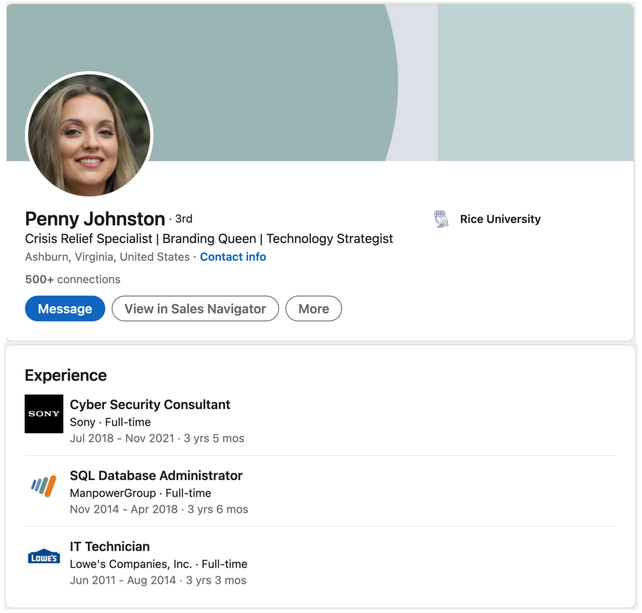
- Why should infosec professionals care about GANs?
- How do GANs work (really short version)?
- What can we do to identify and detect GANs?

Why should I care about GANs?

- You protect a network
- You protect a brand
- You protect users



A fake social media profile created by EXOTIC LILY (Google Threat Analysis Group)



That smiling LinkedIn profile face might be a computer-generated fake

March 27, 2022 | By Shannon Bond

Some of the likely AI-generated faces from fake LinkedIn profiles identified by Stanford University researchers. The central positioning of the eyes is a telltale sign of a computer-created face. (onnie Hanzhang Jin//NPR)

MITRE ATT&CK

T1585.001 - Establish Accounts: Social Media Accounts

ID	Name	Description
G0050	APT32	APT32 has set up Facebook pages in tandem with fake websites.[3]
G0003	Cleaver	Cleaver has created fake LinkedIn profiles that included profile photos, details, and connections.[4]
G0117	Fox Kitten	Fox Kitten has used a Twitter account to communicate with ransomware victims.[5]
G0094	Kimsuky	Kimsuky has created social media accounts to monitor news and security trends as well as potential targets.[6]
G0032	Lazarus Group	Lazarus Group has created new LinkedIn and Twitter accounts to conduct social engineering against potential victims.[7][8][9]
G0065	Leviathan	Leviathan has created new social media accounts for targeting efforts.[10]
G0059	Magic Hound	Magic Hound has created fake LinkedIn and other social media accounts to contact targets and convince them--through messages and voice communications--to open malicious links.[11]
G0034	Sandworm Team	Sandworm Team has established social media accounts to disseminate victim internal-only documents and other sensitive data.[12]

https://attack.mitre.org/techniques/T1585/001/

MITRE ATT&CK

T1585.001 - Establish Accounts: Social Media Accounts

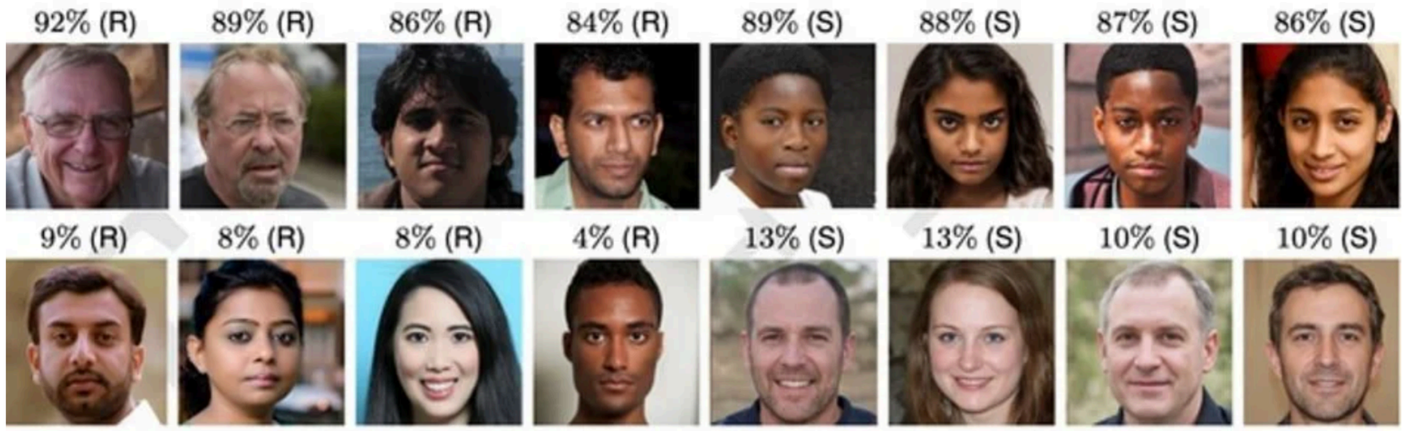
.001	Social Media Accounts	Consider monitoring social media activity related to your organization. Suspicious activity may include personas claiming to work for your organization or recently modified accounts making numerous connection requests to accounts affiliated with your organization.Detection efforts may be focused on related stages of the adversary lifecycle, such as during Initial Access (ex: Spearphishing via Service).

https://attack.mitre.org/techniques/T1585/001/

AI Generated Faces Are More Trustworthy Than Real Faces Say Researchers Who Warn of "Deep Fakes"

A third study asked 223 participants to rate the trustworthiness of 128 faces taken the same set of 800 faces on a scale of 1 (very untrustworthy) to 7 (very trustworthy).

The average rating for synthetic faces was 7.7% MORE trustworthy than the average rating for real faces which is statistically significant.

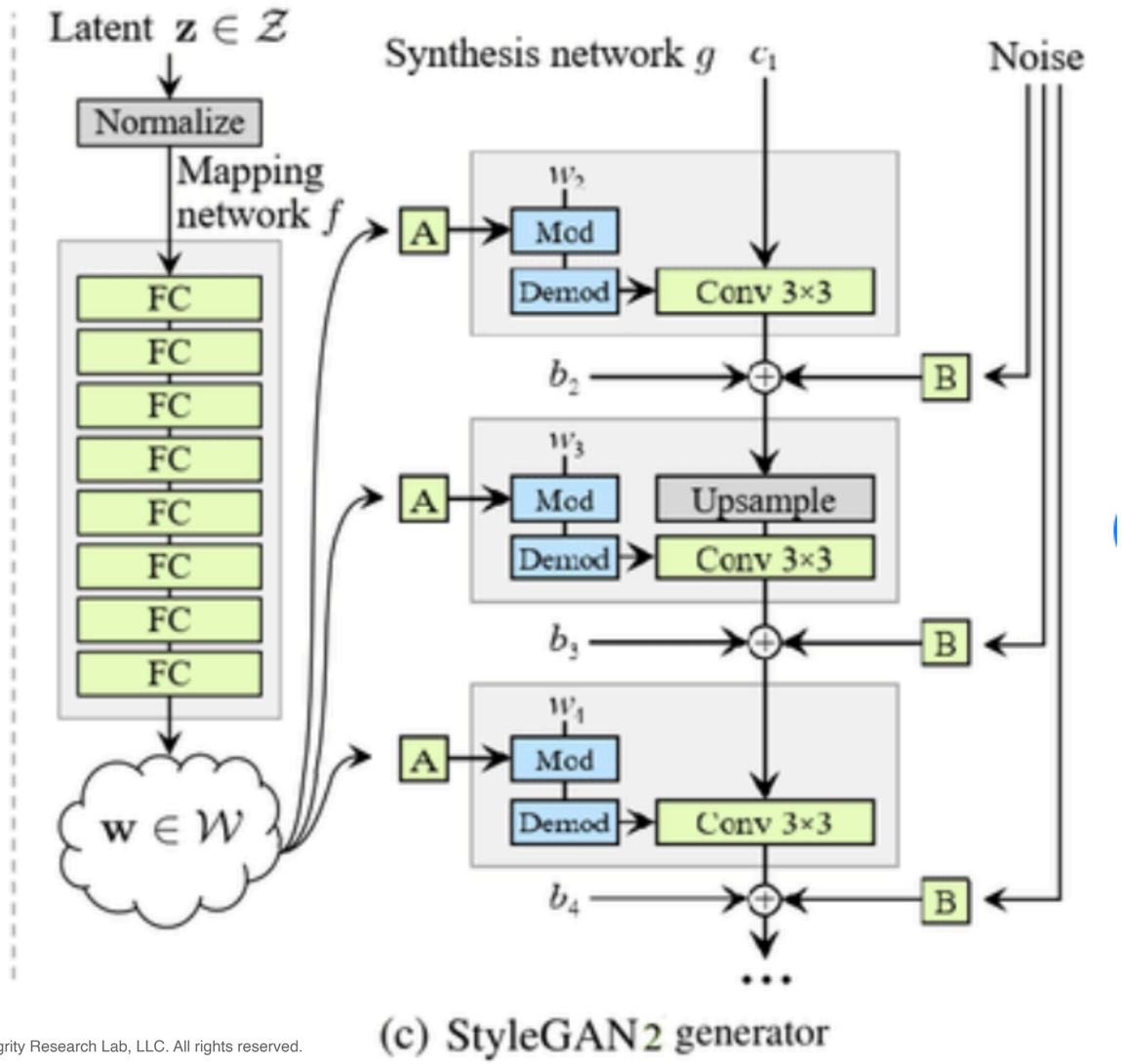


https://neurosciencenews.com/ai-generated-faces-trust-20071/

How do (human face) GANs work?

The details

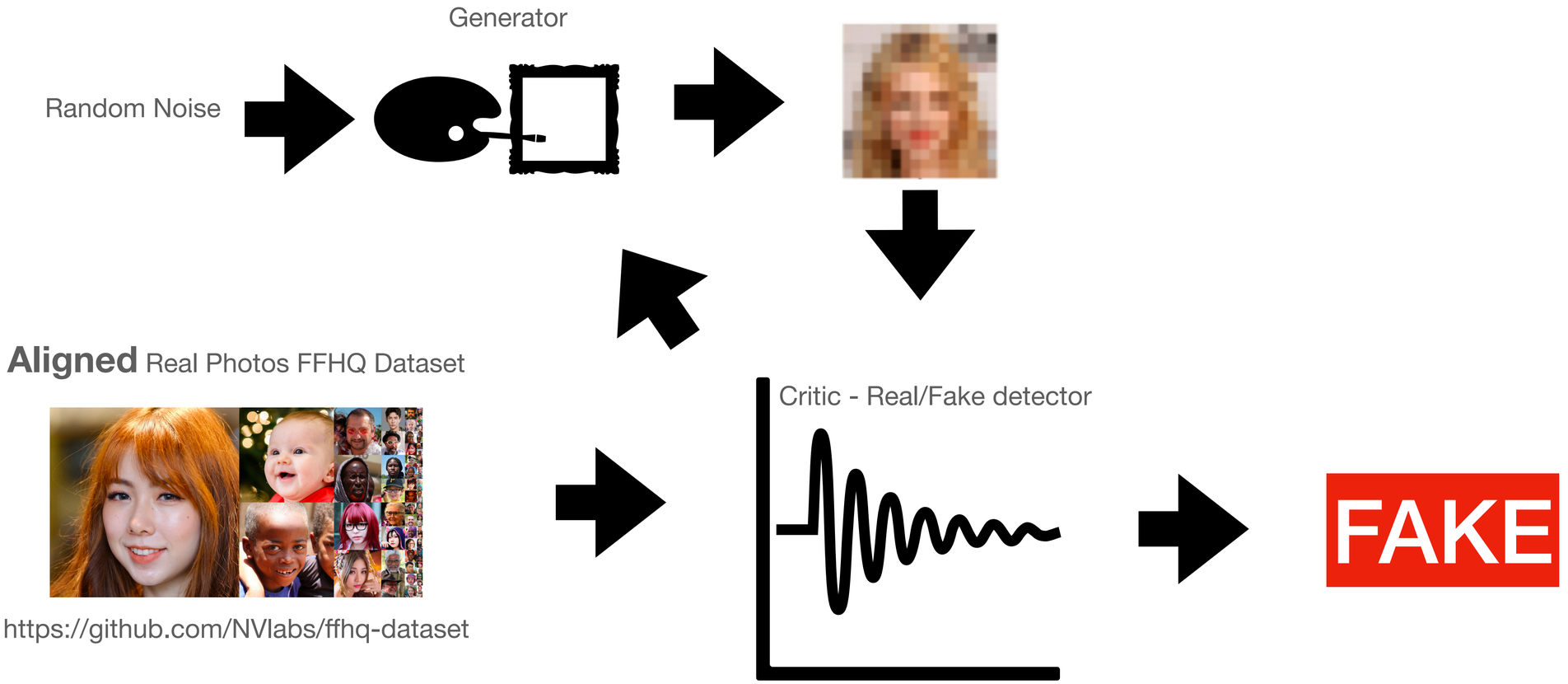
- Generative Adversarial Networks
  - Adversarial competition
  - Generator
  - Discriminator



(c) StyleGAN2 generator

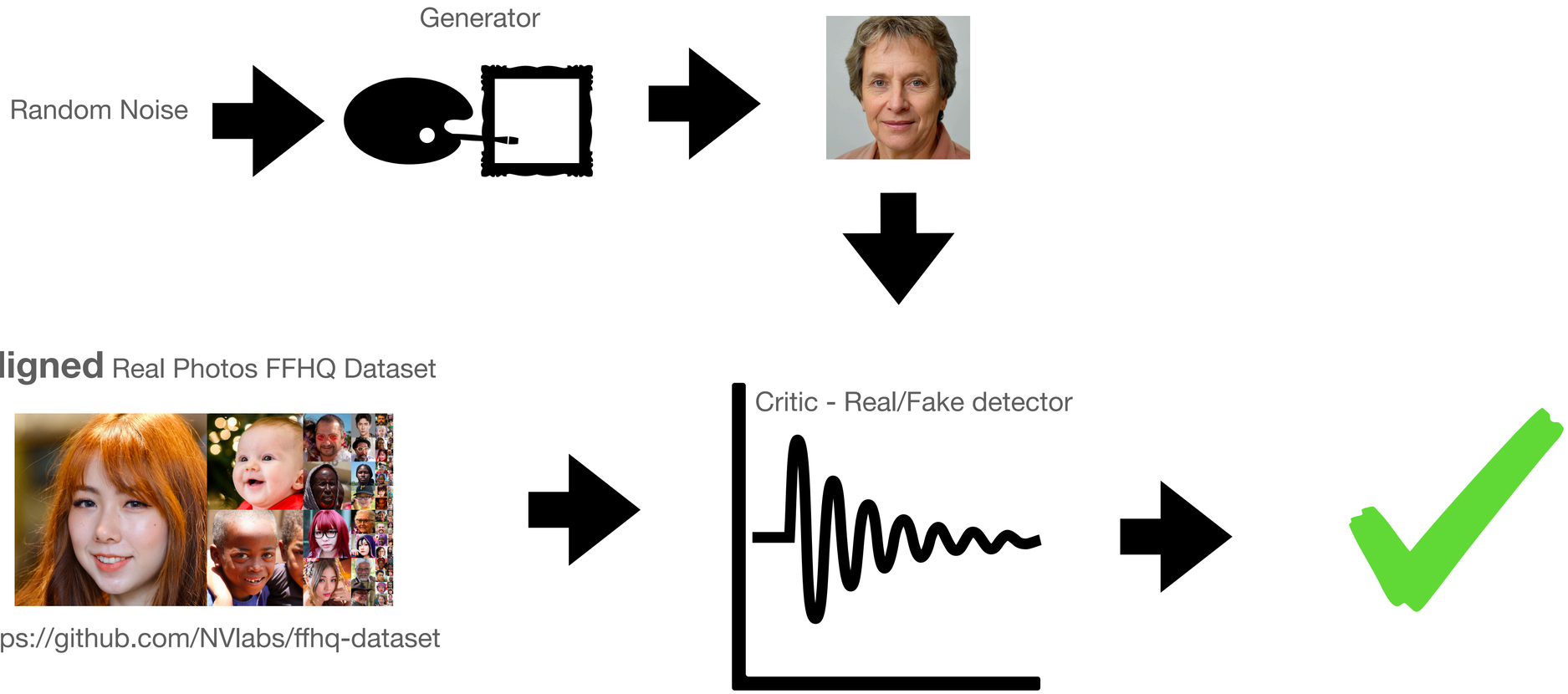
How do (human face) GANs work?

Flickr-Faces-HQ Dataset (FFHQ) -> Linear Algebra -> New Faces



How do (human face) GANs work?

Flickr-Faces-HQ Dataset (FFHQ) -> Linear Algebra -> New Faces

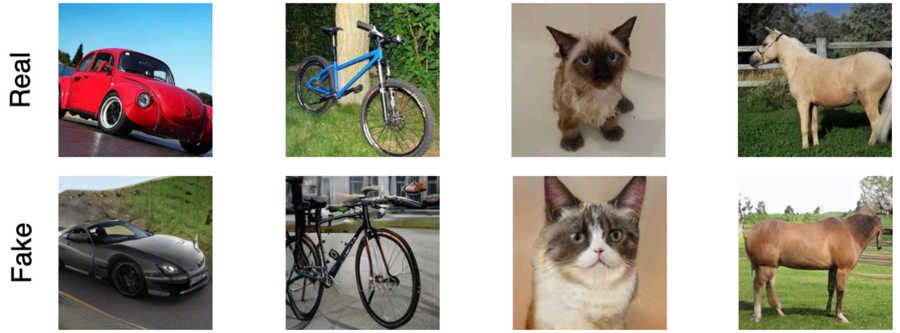


GANs In Wild

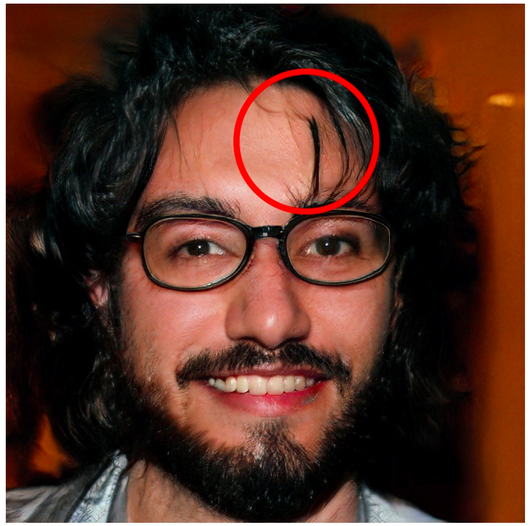
- thispersondoesnotexist.com
- boredhumans.com
- https://github.com/NVlabs/stylegan2
- doesthispersonexist.com/demo
- Cost - 1/100 cent per image
- Low cost + highly scalable + high quality == ripe for abuse

Detection

Existing work



- General NN based techniques for *all* GAN outputs
  - https://github.com/NVlabs/stylegan3-detector
- Visual inspection
  - https://kcimc.medium.com/how-to-recognize-fake-ai-generated-images-4d1f6f9a2842
- “weaponization”



Detection

Exploiting StyleGAN2 structure

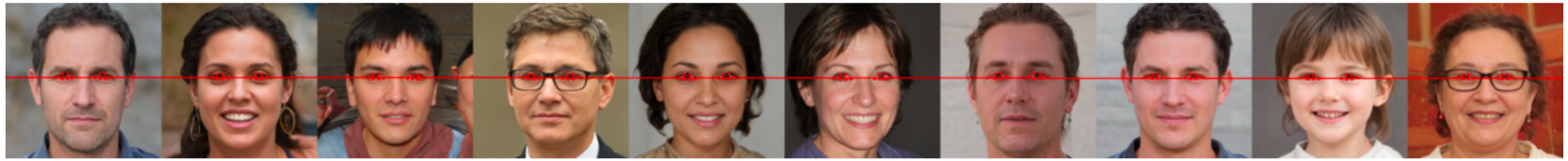
StyleGAN 2 Aligned FFHQ-f

StyleGAN3 Unaligned FFHQ-r

Random Linkedin Profile

StyleGAN2 aligned photos

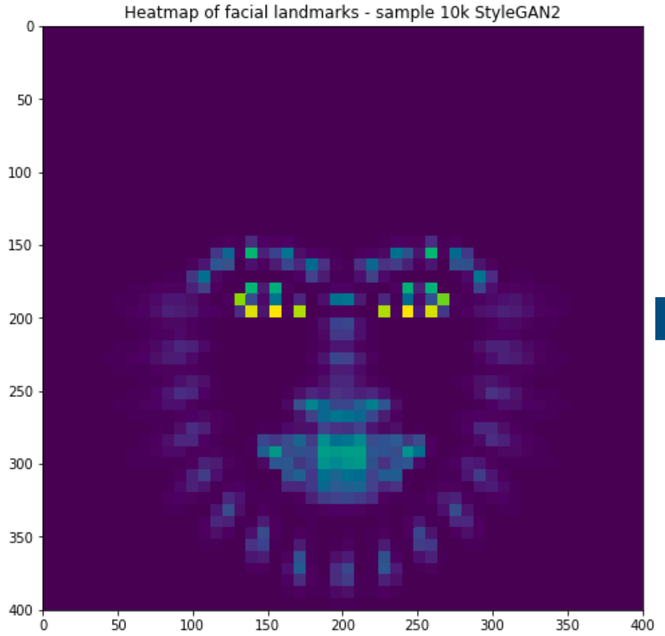
Alignment is a signal (weaponization)



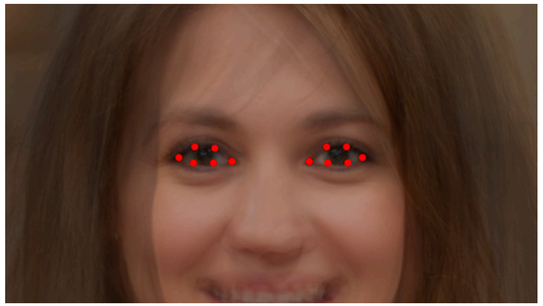
https://medium.com/@mrichard91/pragmatic-detection-of-gan-profiles-f5d7389e40bc

Detection

What artifacts are useful to scale?



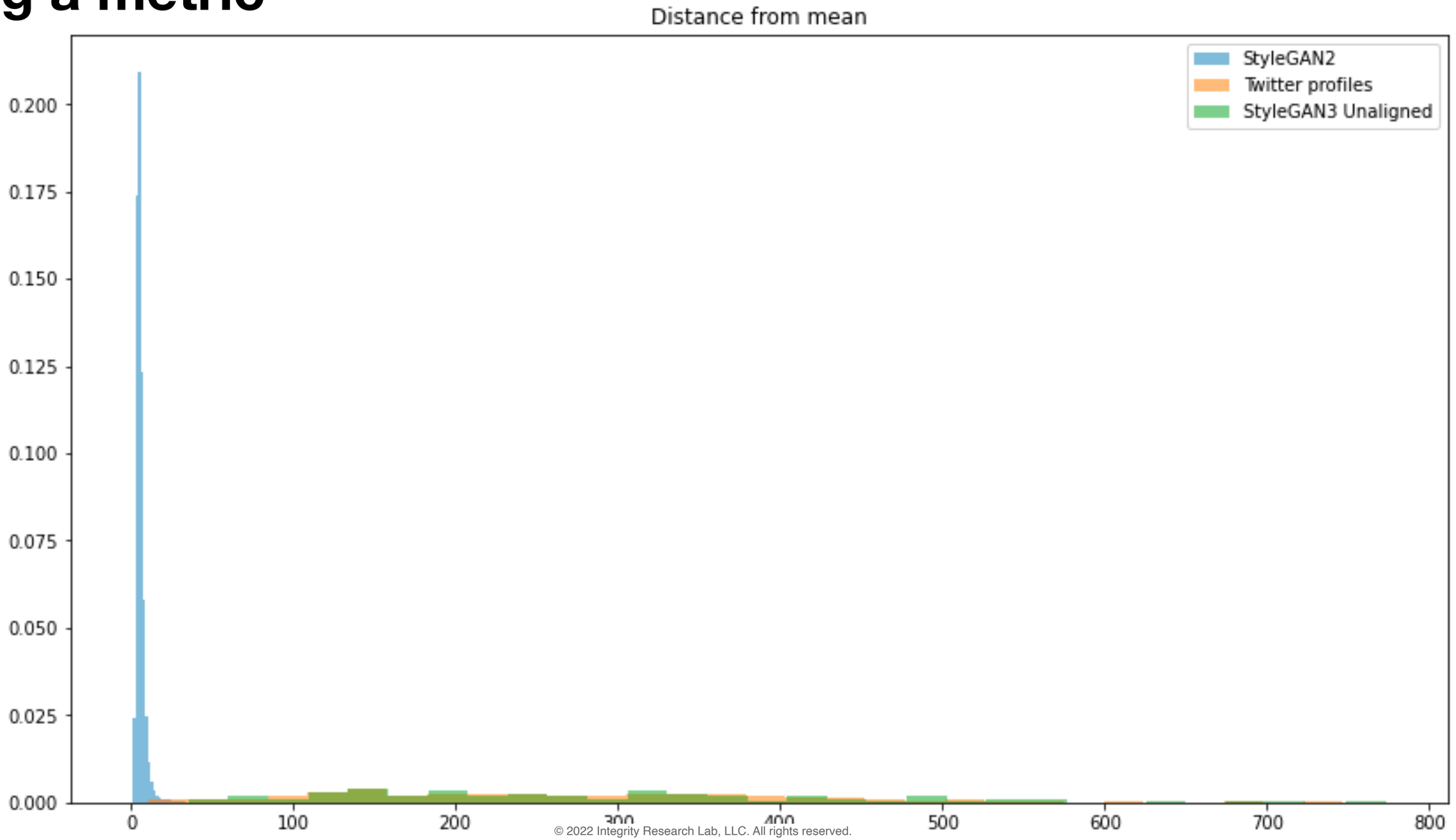
Compute mean eye location



Measure distance from the mean

Detection

Finding a metric



Detection

By the numbers

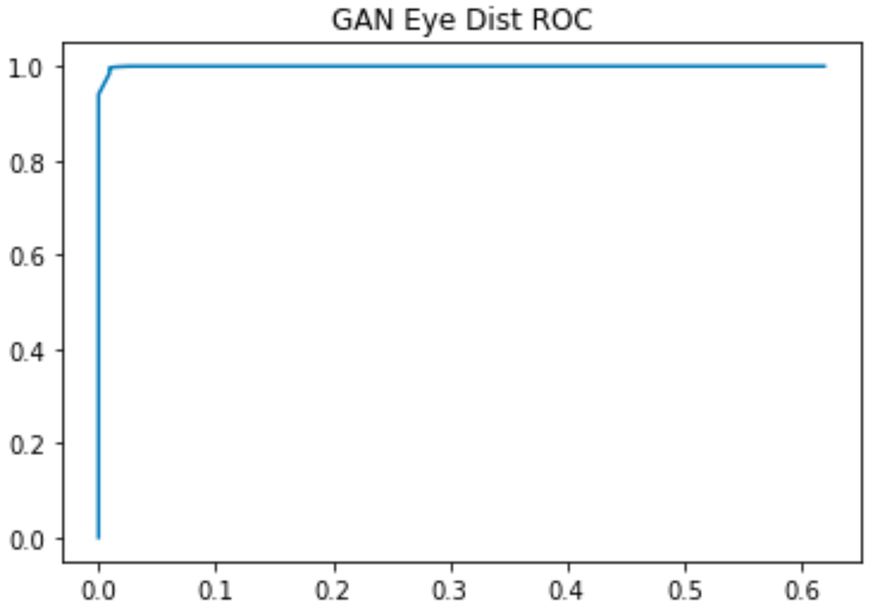
95% of StyleGAN2 < 12 dist

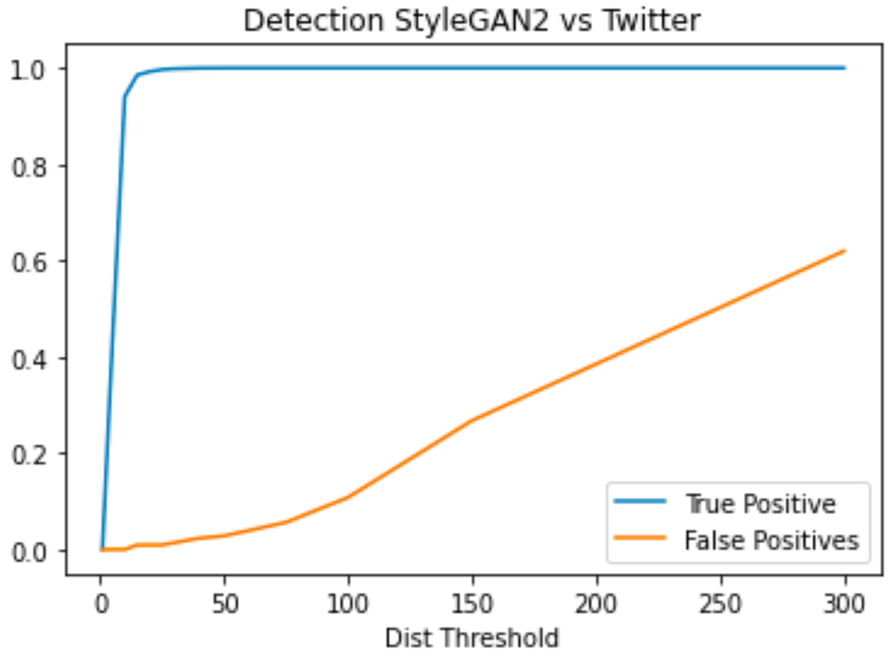
99.7% of StyleGAN2 < 25 dist

100% of StyleGAN2 < 50 dist

99.1% of Twitter > 25 dist

99.5% of Twitter > 12 dist





FP + FN

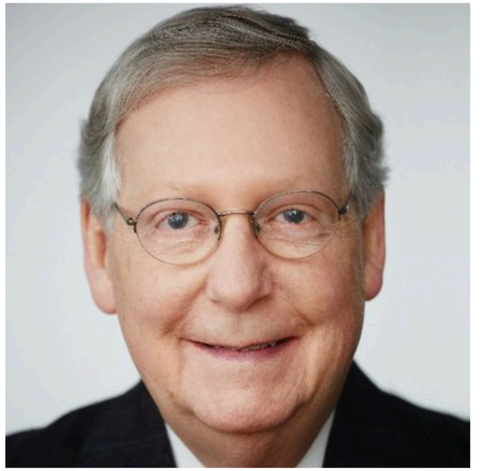
Threshold of 25

False negative due to profile angle

False positive due to intentional alignment

LeaderMcConnell

2.1M Followers



Real Person	★★☆☆☆
StyleGAN2 distance	15.1
Alignment	StyleGAN2 Aligned

Does this person exist (dot com)?

DOESTHISPERSONEXIST DEMO SCAN URL UPLOAD IMAGE

Real Person	★☆☆☆☆
StyleGAN2 distance	10.1
Alignment	StyleGAN2 Aligned
Source	https://thispersondoesnotexist.com/image

Social GAN Scanner support@integrityresearchlab.ai

vs

DOESTHISPERSONEXIST DEMO SCAN URL UPLOAD IMAGE

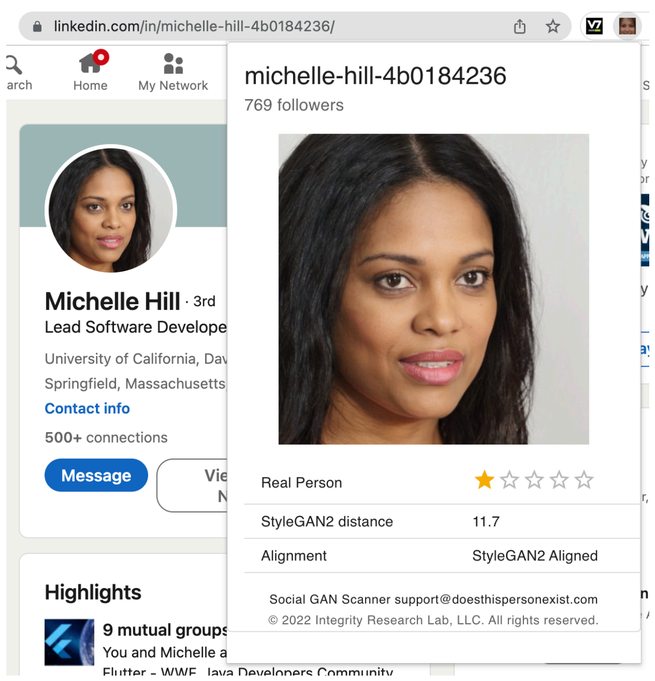
CHOOSE FILES

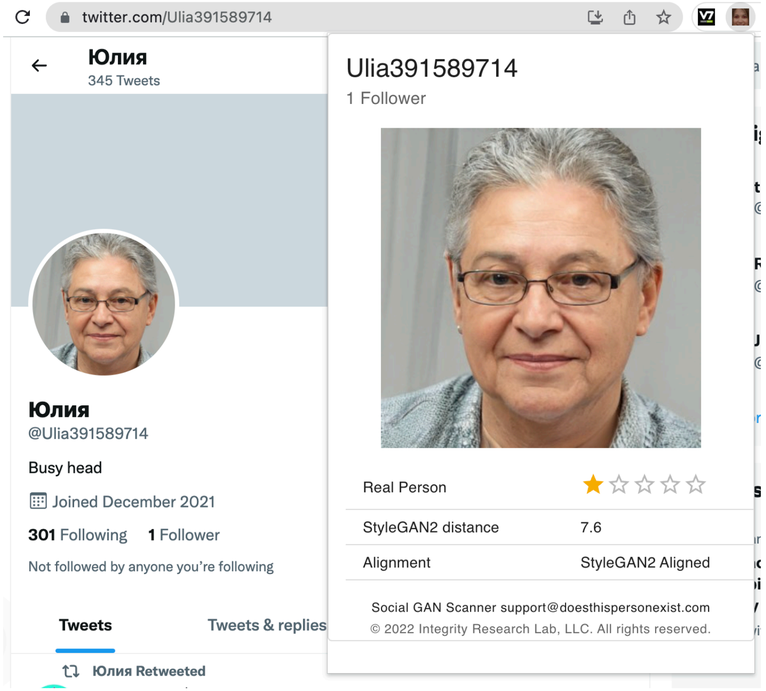
Real Person	★★★★★
StyleGAN2 distance	289.3
Alignment	Not aligned
Source	IMG_1048.JPG

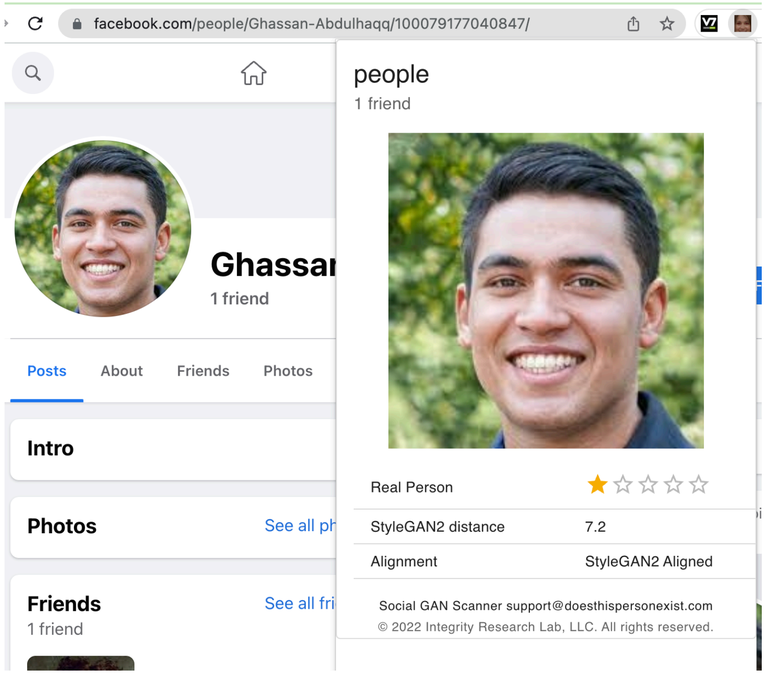
Social GAN Scanner support@integrityresearchlab.ai

https://doesthispersonexist.com/demo

Social Media Triage



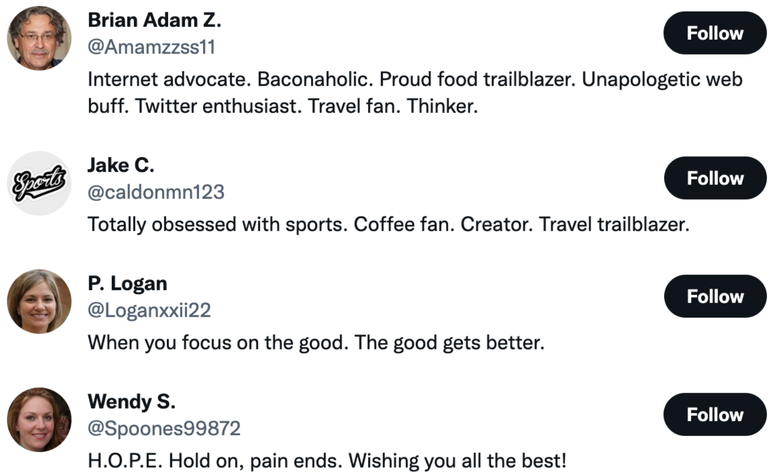


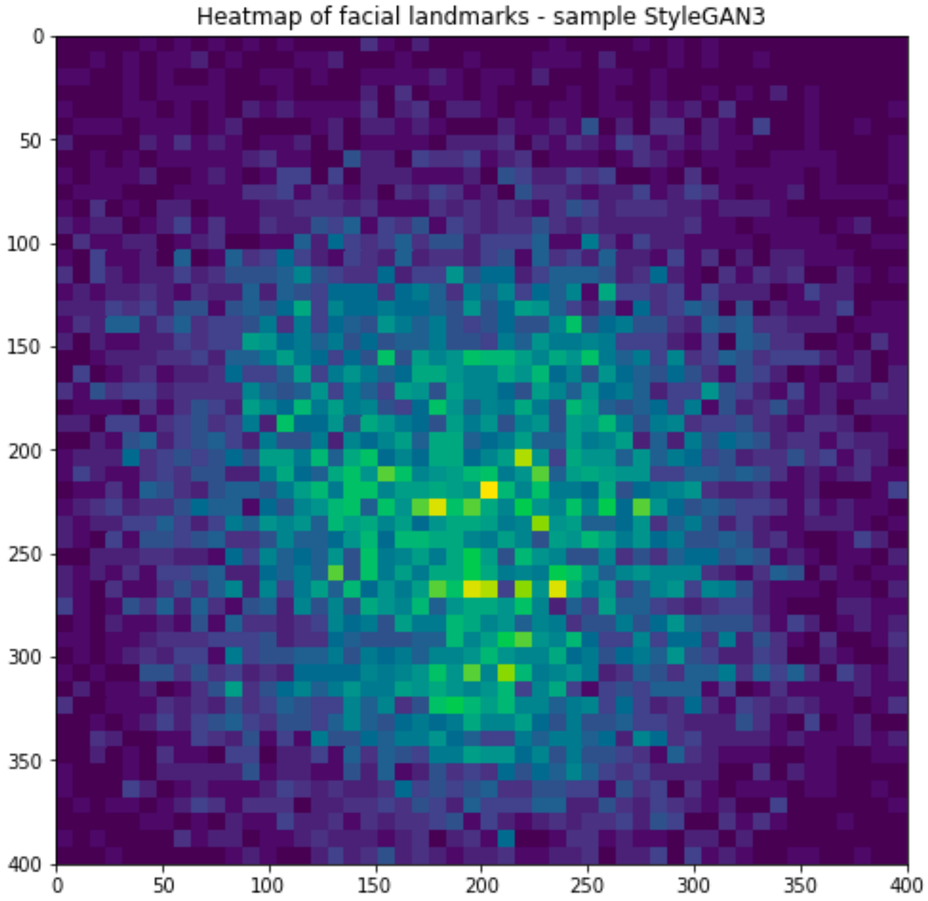


Detection

Future - unaligned models and transformers

- StyleGAN3 has no obvious tool marks
- GPT3 generated text





Other interesting GAN properties

Latent space

Real Me

Me in StyleGAN2 latent space

Aged +25 years

Aged +40 years

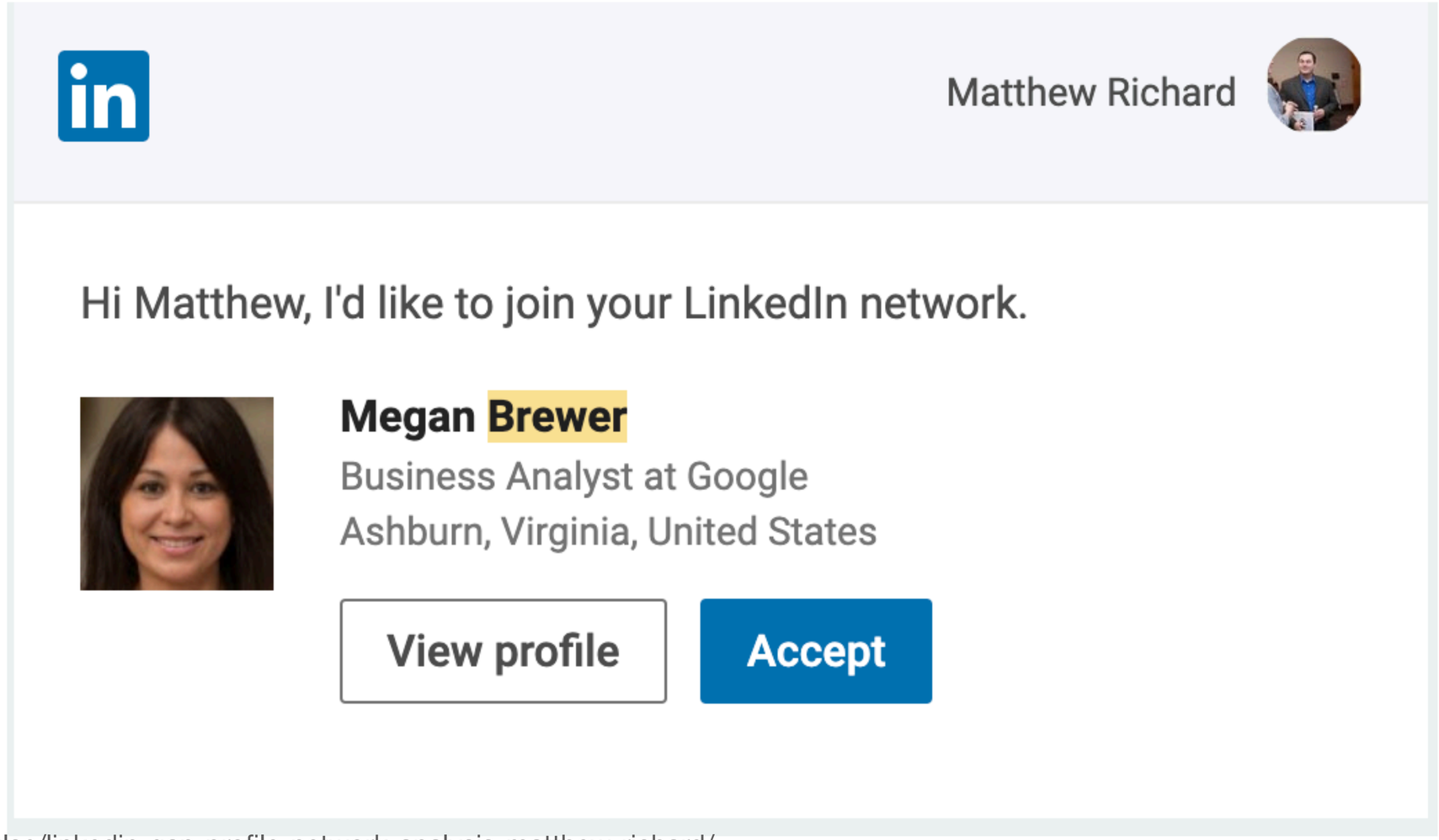
Smiling

Other interesting GAN properties

Infinite possibilities

Detection - LinkedIn Case Study

GAN is a TTP

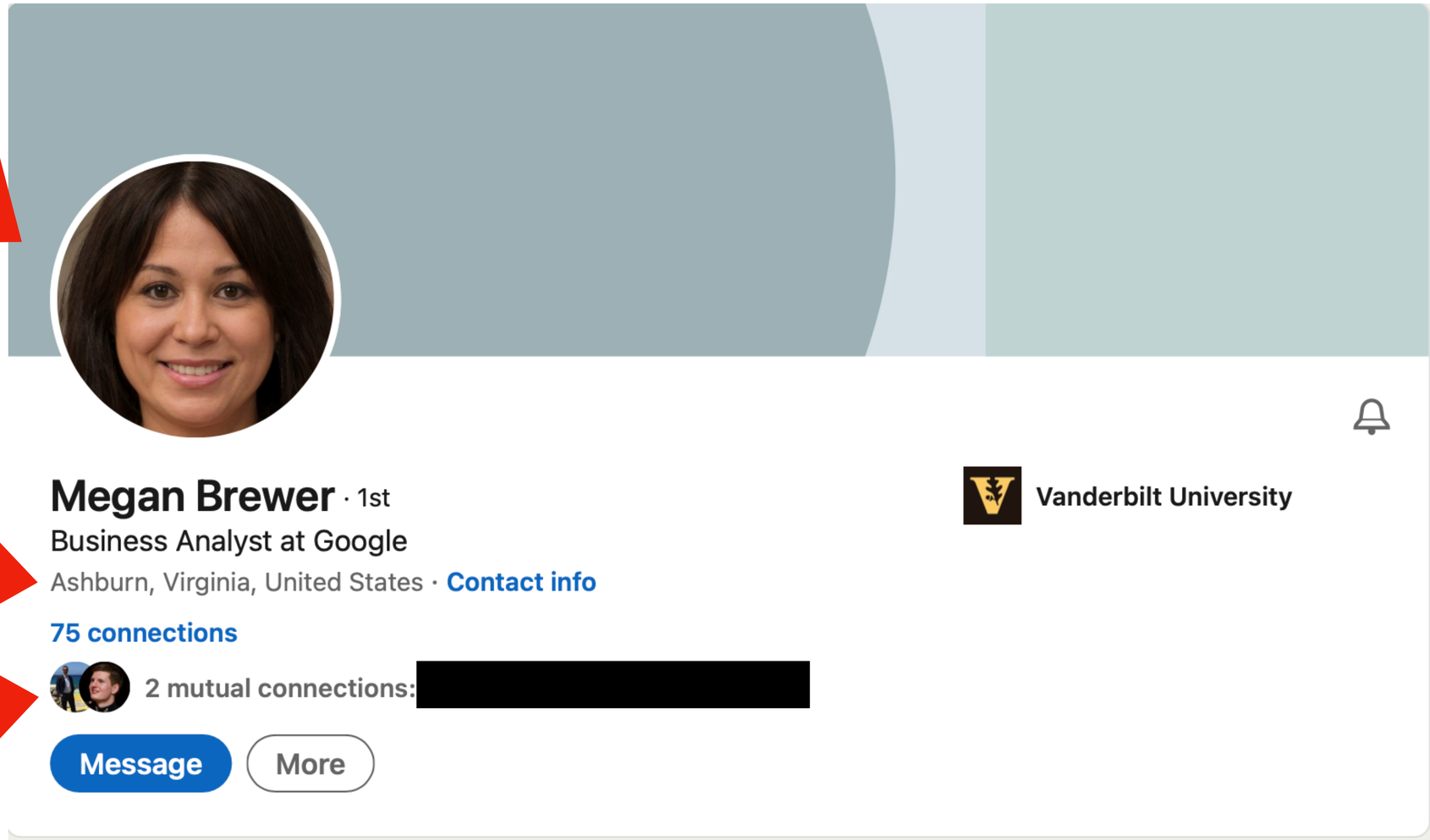


https://www.linkedin.com/pulse/linkedin-gan-profile-network-analysis-matthew-richard/

Detection - LinkedIn Case Study

Profile Analysis

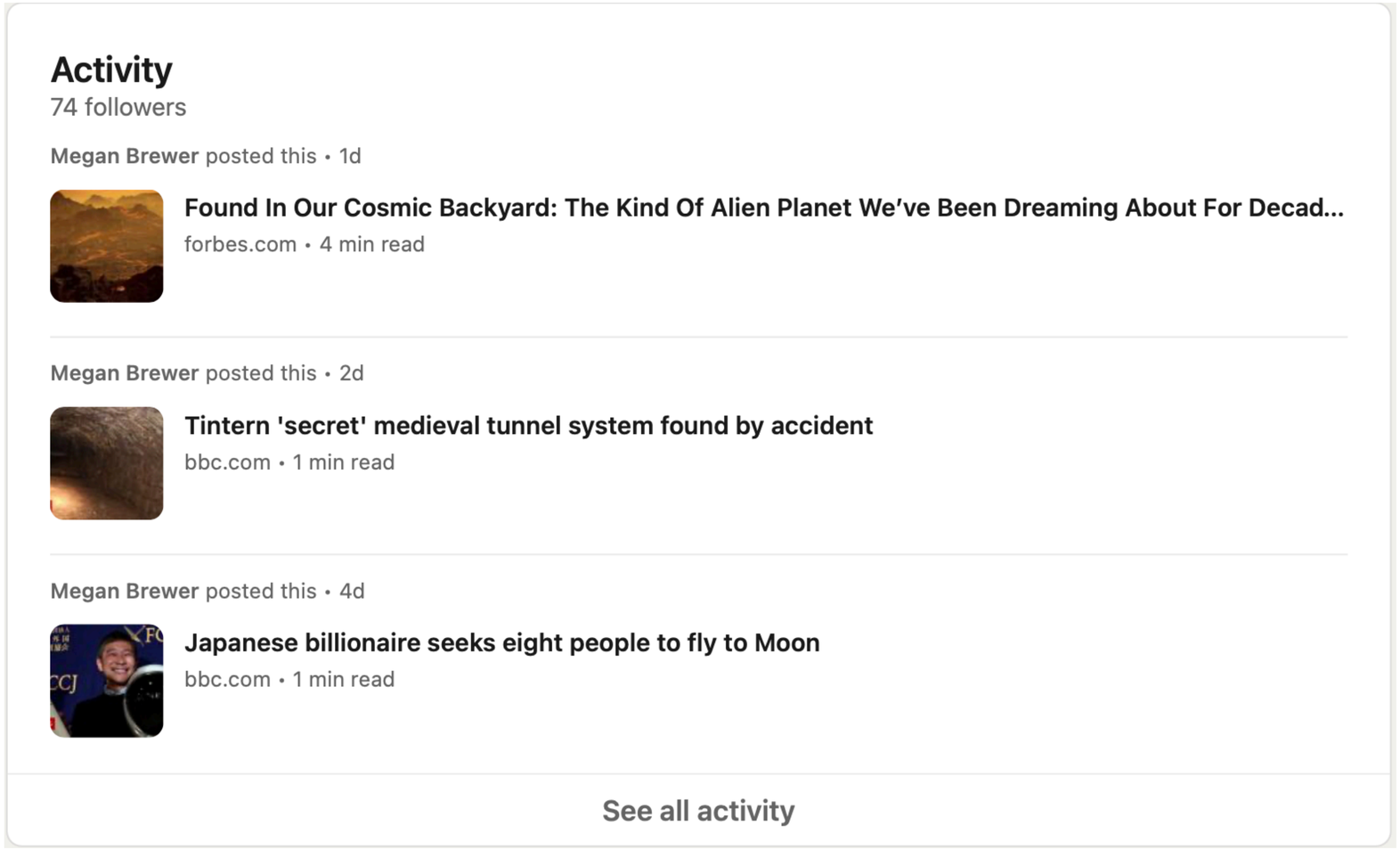
- GAN Photo
- Strange location
- Specific network



Detection - LinkedIn Case Study

Profile Analysis

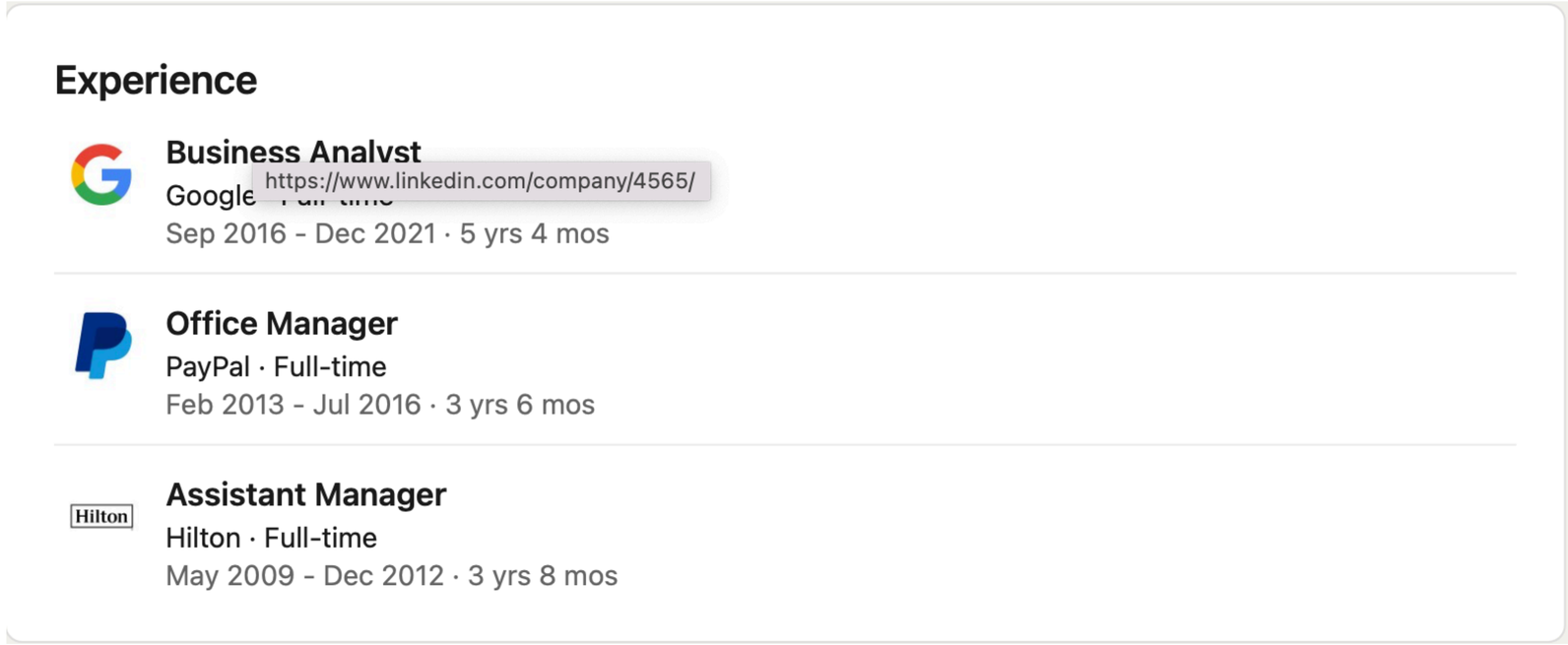
- Random posts



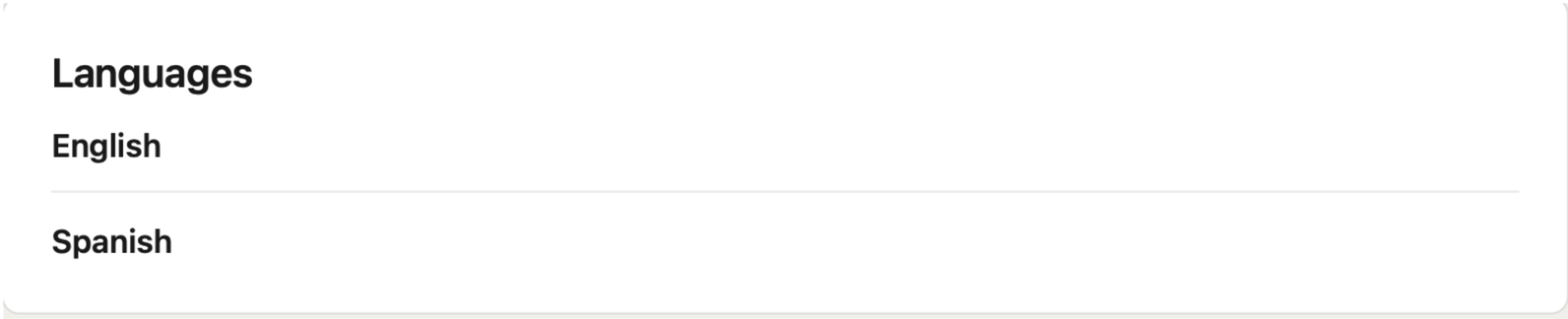
Detection - LinkedIn Case Study

Profile Analysis

- 3 jobs
- 12 years
- Big companies
- Incremental responsibility



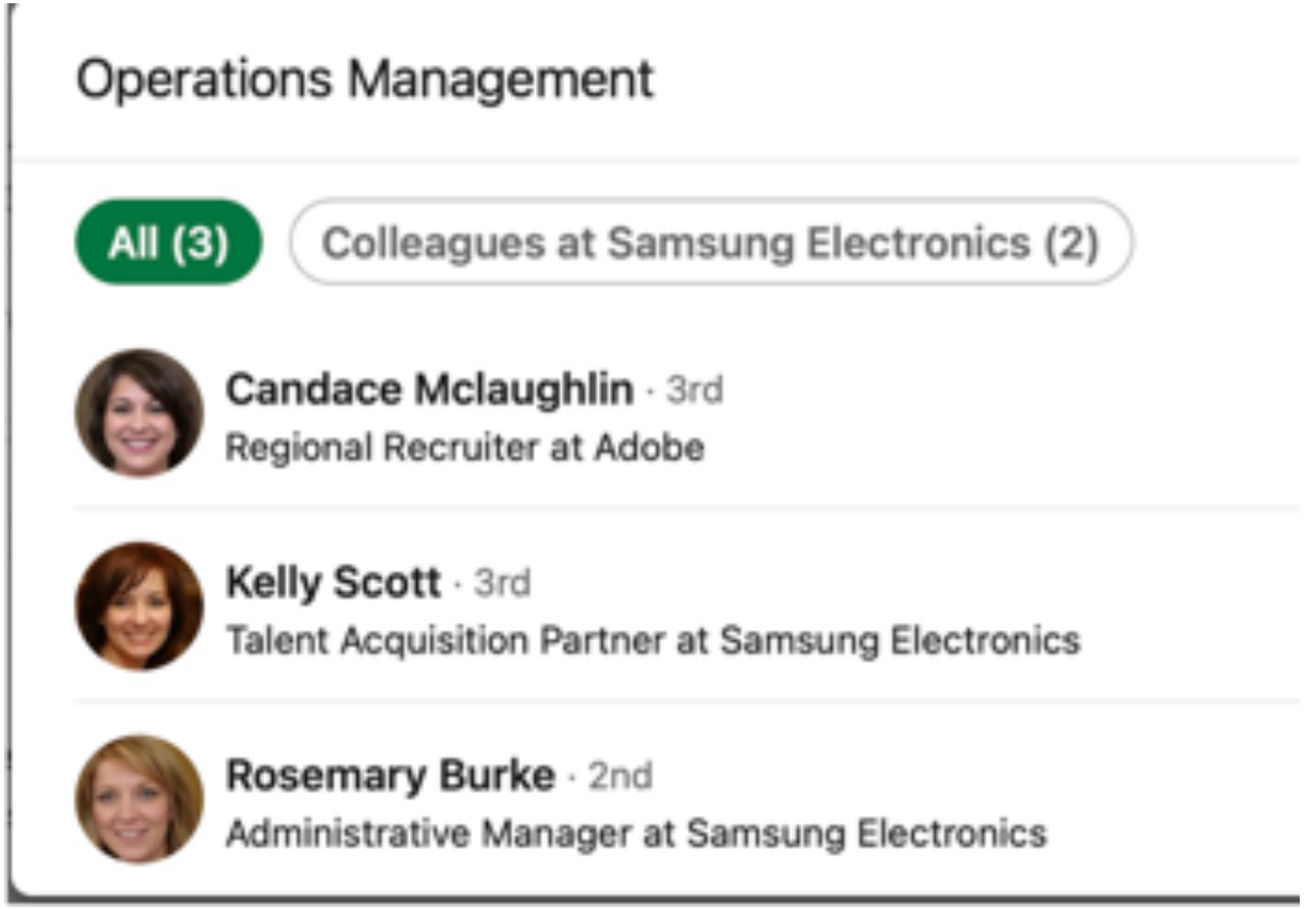
- 2 languages



Detection - LinkedIn Case Study

Profile Analysis

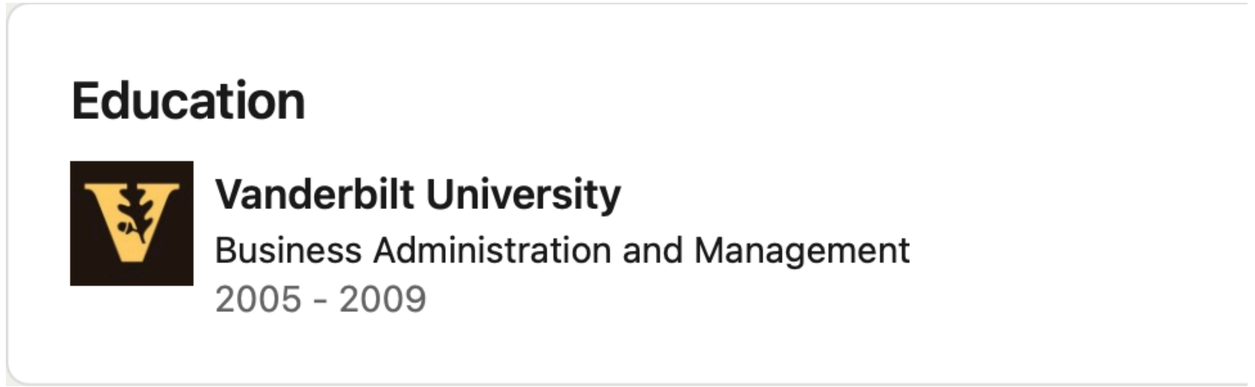
- Endorsements



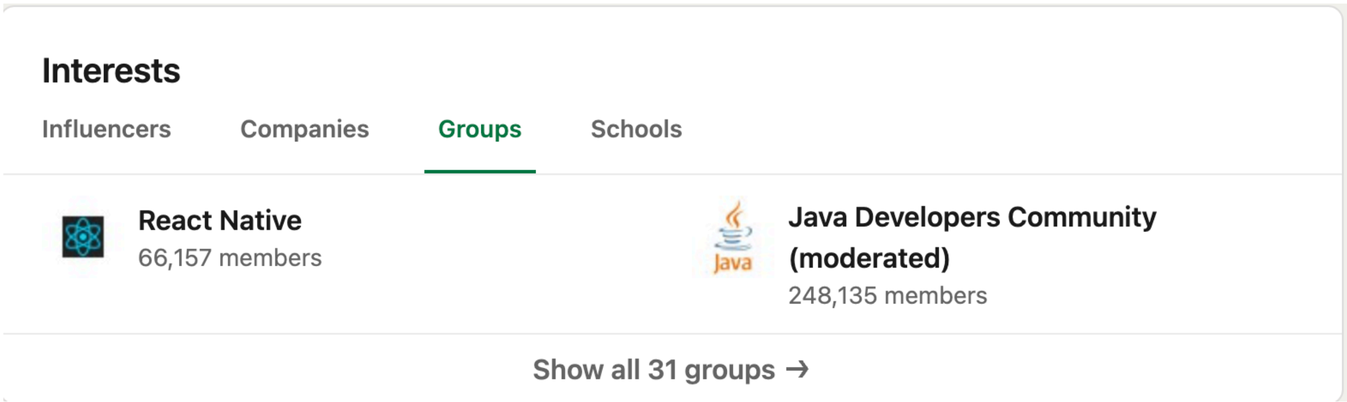
Detection - LinkedIn Case Study

Profile Analysis

- Major US Uni
- Timeline aligned with experience



- Member of many groups
- Groups not matched to jobs



Detection - LinkedIn Case Study

Possible TTPs

- GAN Profile photo
- 3 jobs, incremental, big companies
- US University
- Endorsed by other GAN profiles
- Potentially strange company/location pairs
- 2 languages
- Seemingly random posts
- Joined groups

Detection - LinkedIn Case Study

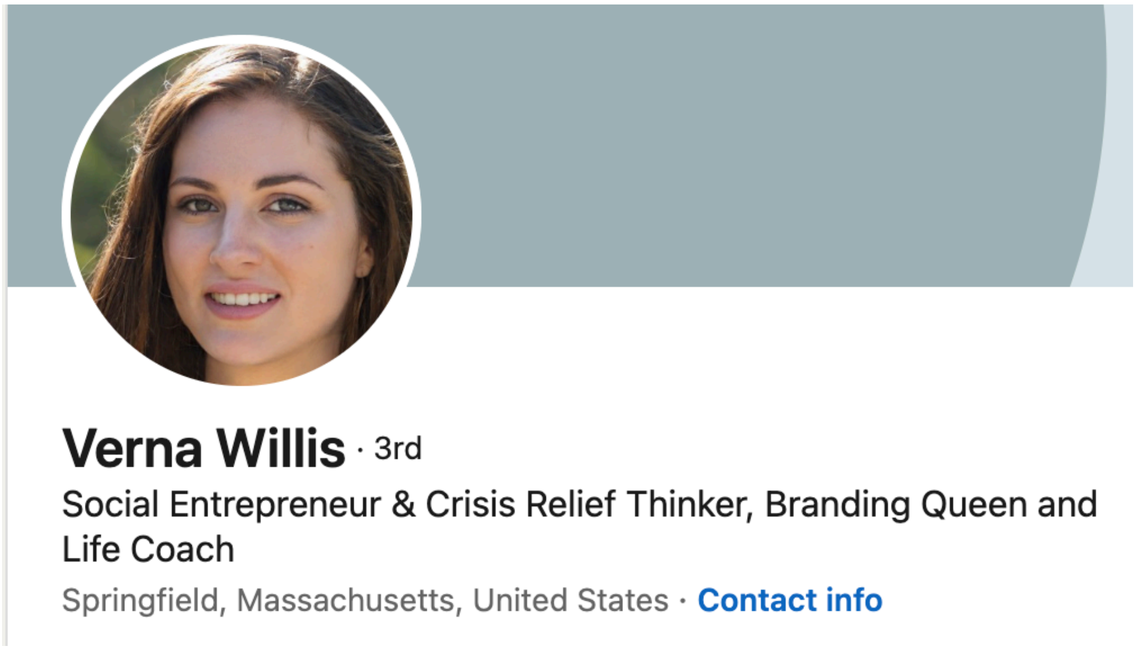
Investigative techniques

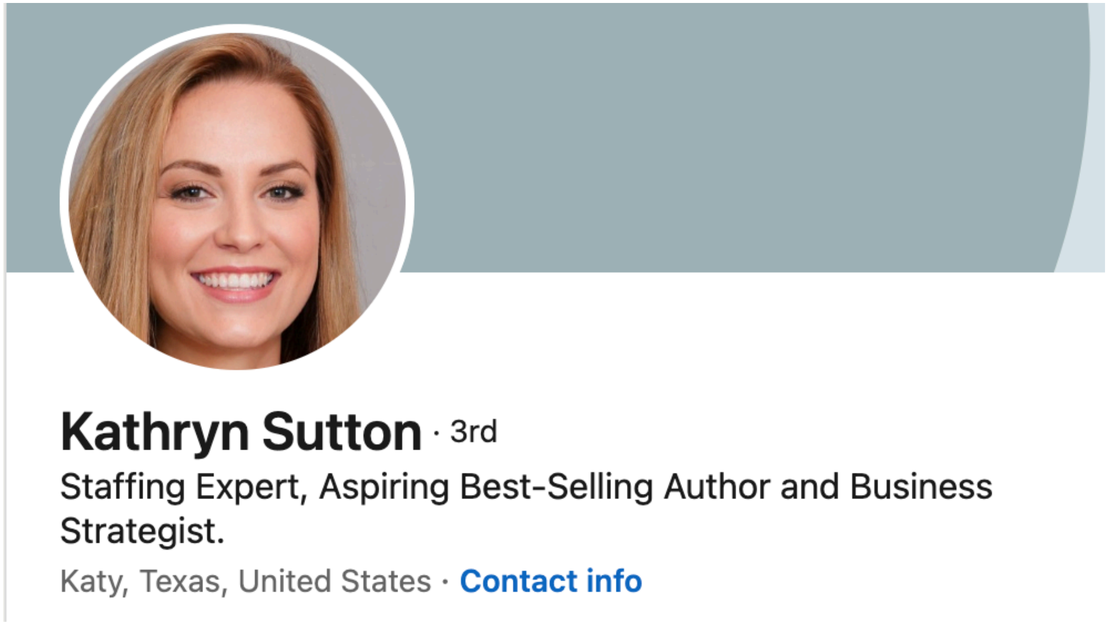
- Tools
  - GAN detector / chrome extension
  - Database/tracker for profiles and data aggregation
- Network fanout
  - Endorsements
  - Group memberships
  - Google dorking
  - LinkedIn Sales subscription

Detection - LinkedIn Case Study

Scaling fake accounts requires automation

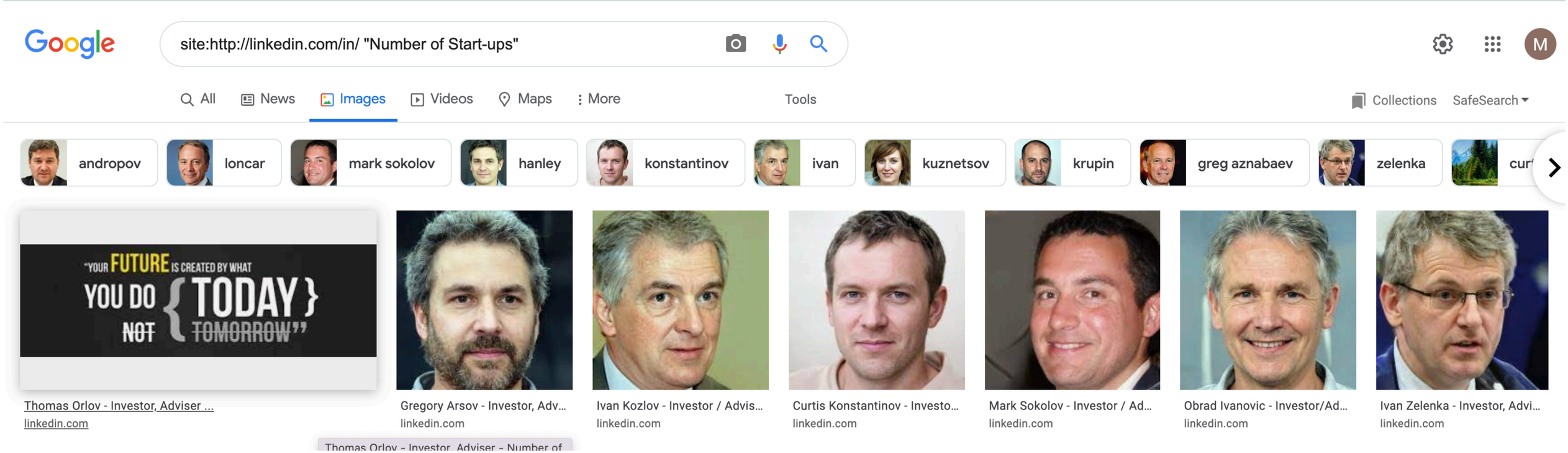
- Generative content - repetitive, improbable





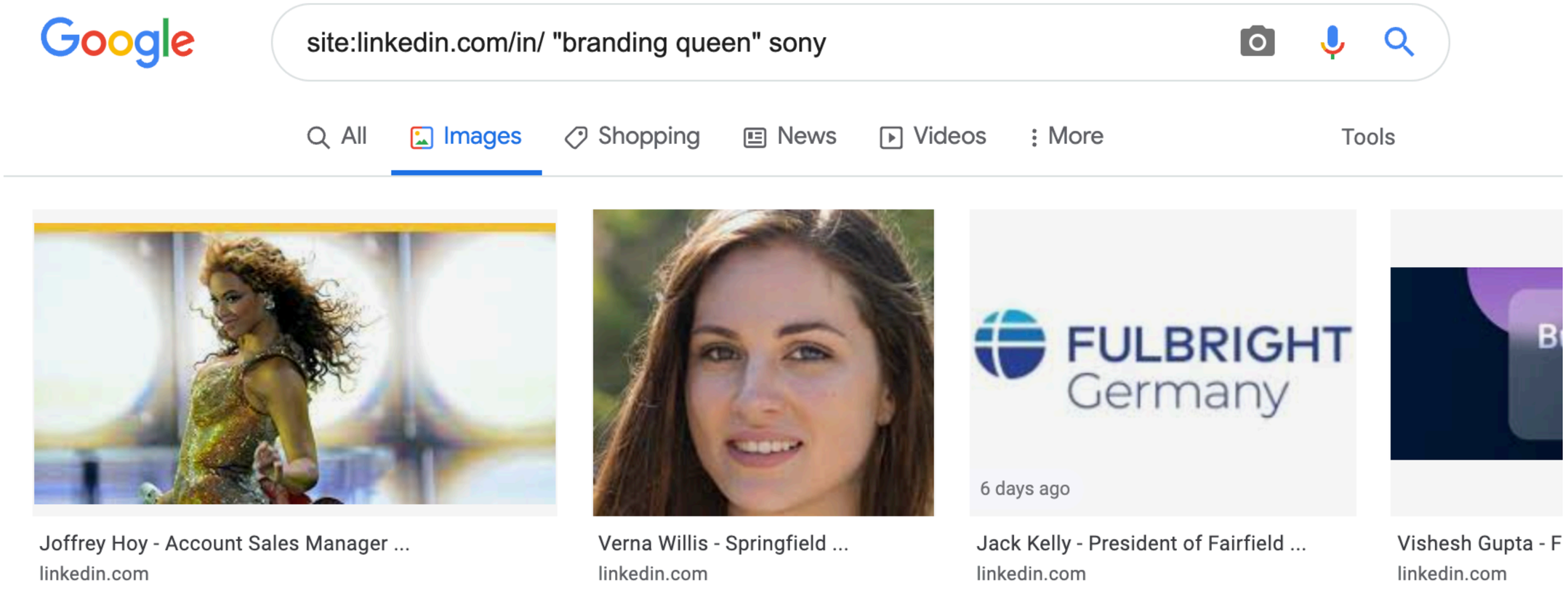
Detection - LinkedIn Case Study

Google dorking



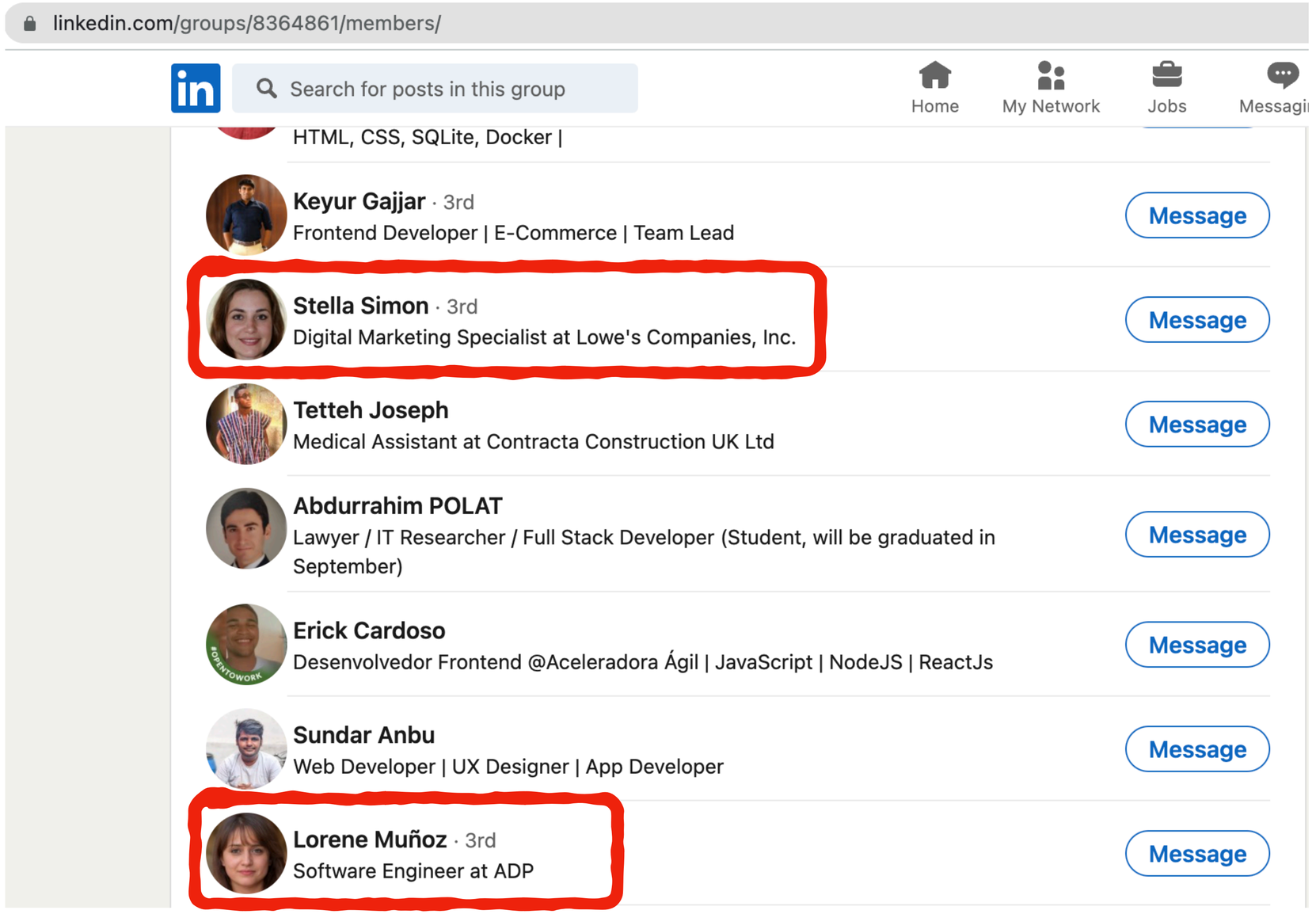
Detection - LinkedIn Case Study

Google dorking



Detection - LinkedIn Case Study

Groups trolling



Detection - LinkedIn Case Study

Where did this go?

- 1,500+ Fake LinkedIn profiles
- 1m+ connections
- 2m+ followers
- 100+ companies

Sony · Full-time	12
Sony Music Entertainment · Full-time	2

Company

ADP · Full-time	73
Western Digital · Full-time	73
PayPal · Full-time	65
Conagra Brands · Full-time	63
Sysco · Full-time	62
Cognizant · Full-time	59
Ford Motor Company · Full-time	54
Arrow Electronics · Full-time	51
Marriott International · Full-time	50
Motorola Solutions · Full-time	50
General Motors · Full-time	49
Nissan Motor Corporation · Full-time	48
Kroger · Full-time	46
Aramark · Full-time	45
Oracle · Full-time	45
Best Buy · Full-time	44
Hewlett Packard Enterprise · Full-time	44

Job titles

Human Resources Manager	118
Human Resources Assistant	88
Media Relations Coordinator	84
Retail Worker	73
Talent Acquisition Partner	62
Human Resources Specialist	60
Human Resources Recruiter	59
Assistant Manager	57
Operations Assistant	56
Recruitment Assistant	54
Junior Software Developer	52
Corporate Recruiter	49
Software Tester	47
Web Designer	43
IT Technician	41
Human Resources Generalist	40
Human Resources Executive	39
Help Desk Analyst	37
Computer Programmer	36
Junior Developer	35

Location

Springfield, Massachusetts, United States	200
Mesa, Arizona, United States	173
Ashburn, Virginia, United States	140
New York, New York, United States	92
Los Angeles, California, United States	67
Montrose, Colorado, United States	61
Charlotte, North Carolina, United States	47
Dallas, Texas, United States	38
Herndon, Virginia, United States	36
Phoenix, Arizona, United States	30
Detroit, Michigan, United States	28
Boca Raton, Florida, United States	28

Tagline

Coding Expert	19
Award-Winning Business Strategist	19
Technology Consultant	19
Crisis Relief Consultant & Expert	17
Expert In Remote Hiring	17
Advocate	17
Technology Enthusiast	15
Technology Expert	15
Advanced	14
Advanced Education Advocate	14
International	14
Technologist	14
Human	13
Social Impact Expert	13
International Crisis Relief Specialist	13

Name

Smith	12
Mary	11
Davis	10
Laura	10
Alice	9
Scott	9
Pamela	9
Melissa	9
Reed	9
Walker	9
Henderson	8
Joyce	8
Lauren	8
Carter	8

Education

2006 - 2010	299
2007 - 2011	271
2005 - 2009	201
Bachelor's degree, Human Resources Management/Personnel Administration, General	116
Bachelor's degree, Computer Science	104

Scaling Detection

Tools + Data + Process + People == winning

- Tools
  - Chrome extension(s)
  - Jupyter notebook
  - Browser automation
  - Detection models
- Data
  - APIs
  - Google
  - Reporting

Key Takeaways

- GANs are a signal of deception
- GANs enable abuse through low cost, high quality, easy to scale content
- Detecting GANs is possible but need to scope
- Consider how to protect using GAN as a signal:
  - Communities
  - Networks
  - Brands
- GANs will only get better, used for more

Interested in more?

- @mjr_irl on twitter
- medium.com/@mrichard91
- linkedin.com/in/richardmatthew/
- matt.richard@integrityresearchlab.ai
- doesthispersonexist.com