Cybersecurity Risks of Social Network Data Aggregation: Leveraging Machine Learning and LLMs in Cloud Environments

Alex Kaplunovich
Department of Computer Science
University of Maryland
Baltimore, MD 21250, USA
akaplun1@umbc.edu

***Abstract*—Social networks have become an integral part of modern life, providing APIs that enable the extraction of vast amounts of user data. Individuals frequently and willingly upload personal information, including photographs, opinions, and geographical locations, across various online platforms. Motivated by the amount of this data, our study aims to quantify the extent of personal information that can be harvested using automated cloud-based serverless architectures, social network APIs, and state-of-the-art Data Science techniques. This paper serves as a compelling exposé on the fragility of digital privacy, demonstrating how easily user data can be aggregated and analyzed through contemporary cloud computing technologies. Utilizing advanced Machine Learning graph models, we extracted a multitude of data points such as geolocations, social connections, similar user profiles, and even made accurate predictions about potential influencers and missing social connections within a user's network.**

Scalable serverless cloud solutions like NoSQL DynamoDB were employed to store aggregated data. Our findings underscore the imperative for individuals to exercise caution in safeguarding their personal information online, as user data can be collected, aggregated, and clustered with ease using modern Generative AI LLMs, RAG and ML techniques. Moreover, our study highlights the risks associated with metadata from camera photos uploaded to social networks. This metadata often includes timestamps, geolocation coordinates, and device information, which can be exploited to track activities, movements, and locations of individuals, effectively turning smartphones into IoT devices that provide continuous data streams. This aspect adds a critical layer to the discussion on cybersecurity, as it exposes how seemingly harmless data can be leveraged for surveillance and profiling. Additionally, we urge social platforms to carefully evaluate the types of user data accessible to third parties to mitigate potential security risks.

Keywords—cybersecurity, serverless cloud, ChatGPT, social networks, NoSQL, LLM, data science, RAG

I. Introduction

Social networking platforms have spread across the digital landscape, each catering to distinct parts of human interaction and serving a myriad of purposes. From professional networking on LinkedIn to creative expression on TikTok, these platforms have become global touchpoints in modern life. They not only serve as conduits for personal and professional engagement but also as lucrative avenues for commercial advertising and user-centric services. Despite the diversification of functionalities—Facebook's foray into dating and live video sharing, for instance—each platform retains a unique focus and a dedicated user base.

The increasing centrality of these platforms in daily life raises critical questions about data privacy and security, topics that have been the subject of prior scholarly investigations ([20], [21], [27]). Moreover, the accessibility of user data through Application Programming Interfaces (APIs) has also been explored ([22], [23], [24]). Our research diverges from existing literature by introducing a novel, trigger-based serverless cloud automation framework that offers scalability, generalization, and systematization, while integrating diverse LLM models, including multimodal ones, for Generative AI inference calls. This methodology is agnostic to the platform in question, enabling the aggregation of data across multiple social networks and the application of deep learning techniques for advanced data analytics.

Each user on these platforms generates a rich tapestry of data, often voluntarily sharing Personal Identifiable Information (PII) such as names, age, email addresses, photographs, and biographical details. Beyond individual profiles, these platforms also maintain intricate networks of social connections—friends, followers, favorites, channels, and invitations—that form the basis of user knowledge graphs. These graphs are not merely digital footprints but valuable datasets that fuel one of the most burgeoning fields in machine learning. They underpin the "friend recommendation" algorithms that many networks deploy to foster new social connections and relationships.

What adds another layer of complexity is the multi-platform behavior exhibited by users. Many individuals maintain presences across multiple platforms, often sharing usernames and social connections, thereby creating a more interconnected web of data. This multi-platform behavior is not just a social phenomenon but also an economic one, as platforms offer monetization opportunities for users with substantial followings. In a bid to capture greater user engagement, platforms are

979-8-3503-5469-0/24/$31.00 ©2024 IEEE

continually refining their algorithms to deliver more personalized and relevant content.

Given the expansive and easily accessible data via APIs, we have developed an automated serverless cloud pipeline designed to harvest a comprehensive set of user information from multiple platforms. Leveraging the capabilities of modern cloud technologies, our system is equipped to handle scalable API requests, offering virtually unlimited storage and computational resources. It also incorporates advanced AI and machine learning algorithms, all while adhering to stringent security protocols.

II. Cloud-Based Serverless Frameworks

In the contemporary landscape of cloud computing, Serverless Lambda functions—or simply "cloud functions"—emerge as a paradigm-shifting service that has revolutionized the way we think about software architecture. This service enables the execution of code at an almost limitless scale, all without the need for dedicated server procurement. This innovation is not just efficient; it is transformative, offering a cost-effective approach to cloud architecture that is unparalleled in today's market.

Each cloud function serves as a discrete microservice, capable of executing specific tasks with a high degree of reliability. These microservices are not isolated; they can invoke other functions and interact seamlessly with various components of your cloud infrastructure. This includes storing results in databases, archiving data in S3 storage, and generating logs for enhanced observability and security measures.

In our implementation, we have leveraged the power of automatic triggering to streamline the operational pipeline, thereby obviating the need for manual intervention. AWS Lambda functions can be triggered by a wide array of AWS service events, ranging from database modifications and storage updates to Alexa commands and messaging activities. Specifically, our pipeline for aggregating user profile information is activated automatically whenever a username or a photo file is deposited into a designated S3 directory.

In summary, Serverless Lambda functions represent a quantum leap in cloud computing technology, offering an agile, cost-effective, and highly scalable solution for modern software development needs. By automating various aspects of the operational workflow, we have been able to focus on what truly matters: delivering exceptional value through superior code quality and innovative features.

A. *Structural design of Microservices*

Microservices represent a transformative approach in software development, as highlighted in various studies [2, 10]. Leading-edge companies often construct their intricate software systems as an ensemble of microservices. These are granular units of functionality that can automatically scale based on demand [19]. Leveraging the platform's APIs allows us to invoke a variety of methods from multiple social networks in order to gather diverse user information. This approach is highly effective due to the inherent design of these services, which are intended to operate autonomously and scale in response to demand. In such a modular setup, each microservice can scale its resources independently, thus offering a more flexible and resilient system architecture.

This level of autonomy and scalability would be virtually unattainable in the context of a monolithic application. In a monolithic architecture, various methods are tightly interwoven and deployed as a single unit, sharing the same server resources. Consequently, these methods are interdependent, leading to limitations in scalability and making isolated updates or modifications a challenging endeavor. The tightly-coupled nature of monolithic applications can introduce resource constraints and make it difficult to adapt to the changing demands or emerging requirements of a dynamic user base.

Therefore, the utilization of microservices not only enhances our ability to gather diverse user data from different social networks efficiently but also provides a more robust and scalable framework compared to traditional monolithic applications.

B. *Architectural Overview*

As illustrated in Fig. 1, our cloud architecture comprises multiple Lambda functions, each designed to perform a specific task that incrementally enriches user data. These functions are employed for a variety of tasks, such as retrieving user details, followers, age, language, and invitee chains.

To overcome Lambda space, code and time limitations, we deploy models into Sagemaker's GPU g4 instances. These models are then invoked from Lambda functions using the boto3 library. This approach enables us to leverage expansive machine learning libraries like Huggingface, Torch, and PyG for tasks such as image similarity, text embedding, and graph and geolocation operations. A salient feature of Sagemaker's inference service is its ability to auto-scale, spawning additional GPU instances in response to a surge in requests.

In summary, our architecture combines the best features of Sagemaker Model Deployment, Inference, and Serverless Architectures to create a robust, scalable, and cost-effective solution. This hybrid approach not only allows us to perform intricate data analyses but also ensures that we remain at the cutting edge of technological advancements in cloud computing and machine learning.

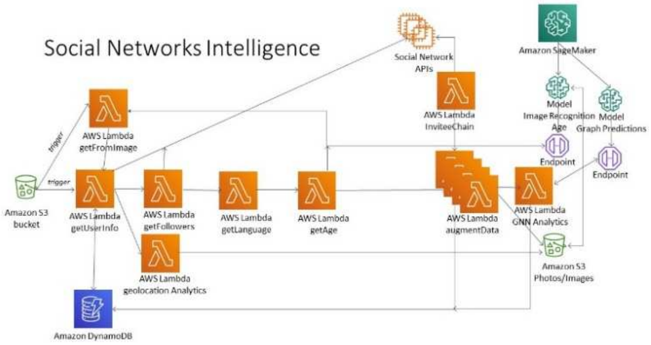


Fig. 1. Architectural diagram

C. *System Configuration and Deployment*

In our study, we observed variations in the type of profile details offered by distinct platforms. Many users are motivated by the desire to amplify their visibility, bolster their popularity, endorse their enterprises, and augment their follower count. Consequently, they often disclose select personal information on public profiles.

The concept of differential privacy has garnered significant attention, as evidenced in [3]. It is evident that a mere public record of one's movie-watching habits can inadvertently reveal sensitive insights, possibly revealing political affiliations or sexual orientations. The depth of personal data on social networks is even more profound, encompassing actual names, perspectives, writings, and photographs, often readily available via public APIs.

We have executed a suite of foundational operations through Lambda functions that engage with public methods of social networks:

1) Profile Retrieval: Numerous APIs facilitate user queries based on name, pseudonym, or specific interests (like soccer or films). The resultant data might encompass biographical details, user IDs, URLs of photographs, as well as tallies of friends and followers. It's noteworthy that several social media profiles embed hyperlinks directing to the user's profiles on alternate platforms - such as Instagram, LinkedIn, or Twitter. This interconnectedness can significantly enrich our understanding of an individual.

2) Follower Acquisition: This method is pivotal for observing user preferences, given the tendency of like-minded individuals to converge.

3) Invited Profile Retrieval: Some platforms maintain a log of users one has invited, facilitating the construction of a connectivity chain to the originating user.

4) Photograph Collection: Contemporary photographs are data troves, often embedded with geolocation data and timestamps. Additionally, facial recognition technology can streamline user identification.

5) Influencer Identification: Profiles with an extensive follower base invariably command attention, generating heightened activity. These individuals, frequently dubbed "influencers", are of paramount importance.

Beyond the mere collection of profile data, our system can perform intricate data analytics, deploying a diverse array of machine learning paradigms and methodologies. These data science techniques, harnessed by our serverless functions, enable profound data comprehension and information augmentation. While machine learning capabilities in Natural Language Processing (NLP) and vision are considerably advanced, the judicious selection of a fitting model is imperative to ensure both precision and cost-effectiveness.

D. Features Implementation

1) Language Detection

For various data science tasks, there is often no necessity to train new models from scratch. We have employed the LangDetect library, a Google original, which is proficient in identifying up to 55 languages [4] (as illustrated in Fig. 2). Interestingly, LangDetect operates seamlessly without the need for GPU acceleration and boasts an impressive accuracy rate of up to 98% [5]. This feature aligns ideally with our objectives. The algorithm exhibits remarkable speed: it can identify the language of 60 texts per second when deployed on a serverless Lambda function, translating to a processing capacity of 216,000 documents per hour. The library effectively covers a global linguistic landscape.

af, ar, bg, bn, ca, cs, cy, da, de, el, en, es, et, fa, fi, fr, gu, he,hi, hr, hu, id, it, ja, kn, ko, lt, lv, mk, ml, mr, ne, nl, no, pa, pl,pt, ro, ru, sk, sl, so, sq, sv, sw, ta, te, th, tl, tr, uk, ur, vi, zh-cn, zh-tw

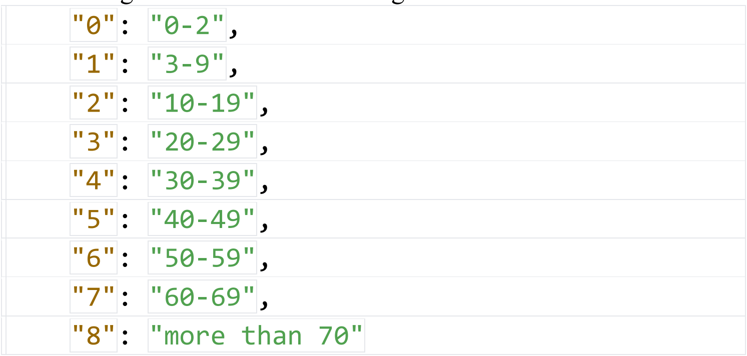
Fig. 2. Languages supported by langdetect

2) Phography visualization

An overwhelming majority—more than 75% of users—have biographies. Automatic visualization of these biographies to elucidate entity relationships is crucial. Among multiple approaches to text relation discovery, a relatively simple yet effective technique is using object-verb-subject triplets via Textacy [6]. This strategy is both quick and does not necessitate the use of resource-intensive GPU instances. Alternatively, the LUKE model from Huggingface [7] can identify up to 41 textual relations, offering the possibility of fine-tuning if labeled data is available.

3) Photographic Age Detection

Age determination from photographs is a standard computer vision task. We employed the vit-age-classifier model from Huggingface, which outperforms traditional methods on the ImageNet dataset [8, 9]. Due to Lambda's resource limitations for GPU-accelerated torch models, deployment on Sagemaker became necessary. The pretrained model classifies age into various ranges as demonstrated in Fig. 3.



```
    "0": "0-2",
    "1": "3-9",
    "2": "10-19",
    "3": "20-29",
    "4": "30-39",
    "5": "40-49",
    "6": "50-59",
    "7": "60-69",
    "8": "more than 70"
```

Fig. 3. Image age classification ranges

These categories facilitate approximate age identification, aiding in user grouping. Using a similar technique, we can identify a gender.

4) Text Embedding for Profile Similarity

State-of-the-art NLP technologies can transform text into numerical vectors, enabling similarity assessments through cosine distance or other metrics. The BERT sentence transformers [12] currently offer the most effective results for document embedding [11]. This machine learning method can uncover users with similar biographies and extend the analysis to posts, comments, and messages.

5) Identity Validation Across Photographs

One crucial vision task involves recognizing the same individuals across multiple photographs. To achieve optimal results, we employ leading-edge vision neural networks, specifically FaceNet from Google [13] and face-compare [14]. Our experimental results corroborated the paper's [13] claim of over 95% accuracy in identity validation.

6) Knowledge Graphs for Relationship and Profile Classification

Social networks essentially operate as directed graphs, connecting followers and friends. Using this structure in conjunction with biographical data allows for friend suggestions. We have also explored invitation chains to trace the origin of each connection, utilizing PyTorch Geometric graph networks [15] and geometric deep learning [16] for efficiency. For graph visualization, we employed the NetworkX library [17], which offers comprehensive functionality for graph representation.

7) User clustering and bio summarization with LLMs like ChatGPT

Modern Large Language Models can understand user texts easily. Moreover, we can ask the model like ChatGPT or Claude to summarize multi-lingual profile bio. Asking a simple question like “Identify top ten subjects (in English) from text” will give up to ten user topics, and “Summarize the bio below in up to two English sentences” will generate a concise summary. That information can help us to cluster users and condense her message in one specified language.

It is important to mention that we can utilize commercial models like OpenAI’s GPT-4o as well as comparable open source models (Llama3 or Mixtral). Although it is beyond the scope of the paper to discuss detailed advantages and disadvantages of LLM types. However, we would like to mention that open source models supported by Huggingface common interface give data scientists full control of deployment, architecture, finetuning, cost and model selection. Additionally, utilization of llama_cpp library [25] can help scientists to avoid high costs associated with expensive GPU instances and library dependencies.

Through these multiple AI avenues of analysis and data processing, our system architecture offers a holistic approach to understanding user behavior and relationships in social networks.

III. Results

Ascertaining information about user profiles is a considerably straightforward endeavor. Our analysis unveiled an array of compelling data points, encompassing user statistics, intricate details, multimedia content, and other essential indicators. This resonates with the well-established saying, "Tell me who your friends are, and I'll tell you who you are." The saying is particularly useful in the context of social networking platforms. Friends and followers serve as distinctive markers, elucidating individual preferences and character traits with considerable accuracy.

When data from multiple social media platforms is integrated, the potential for precision in user profiling is amplified exponentially. Beyond the utilization of advanced graph analytics or Natural Language Processing (NLP) techniques, profiles that share a high number of common connections are probabilistically more likely to engage with one another.

In the realm of social networks, connections are not merely digital linkages but proxies for a myriad of behavioral and psychological factors. Consequently, understanding these connections can yield substantive insights into user behavior and preferences. Thus, in our research, we have ventured beyond traditional graph and NLP algorithms to create a multi-layered framework that can more holistically analyze and predict potential connections based on shared friendships and follower relationships.

This multi-sourced data augmentation not only enhances the richness of our profiling but also introduces new avenues for predictive analytics, enabling a more nuanced approach to recommending potential connections or friendships on social platforms. Therefore, our methodology represents a comprehensive approach to comprehending the complex fabric of relationships and user behavior in digital social ecosystems.

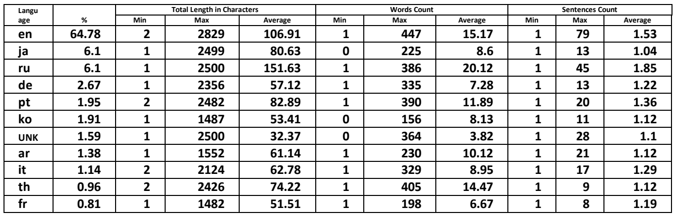
A. Cumulative users results

Our research has yielded compelling statistics concerning the distribution of various attributes within our user base. These attributes encompass aspects such as friend count, follower count, invitee chain length, age distribution, and biography length. Fig. 4 and Fig. 5 provide detailed graphical representations of these distributions, offering nuanced insights into user behavior and demographics.

A noteworthy observation is that a majority of profiles possess fewer than 250 followers. Specifically, the influencer subset—profiles with more than 1,000 followers—comprises a minuscule percentage of the user base, estimated at less than 3.6%. This illuminates the disproportionate influence a select few wield over the larger community, thereby highlighting the skewed nature of follower distribution in social networks.

Another intriguing data point is the age distribution among users; a staggering 92% fall within the 20-40 age bracket. This demographic information is particularly significant as it indicates that the platform predominantly appeals to a relatively youthful audience, who are often more engaged and active on social media platforms.

TABLE I. PROFILE BIO STATISTICS



Language	%	Total Length in Characters			Words Count			Sentences Count		
		Min	Max	Average	Min	Max	Average	Min	Max	Average
en	64.78	2	2829	106.91	1	447	15.17	1	79	1.53
ja	6.1	1	2499	80.63	0	225	8.6	1	13	1.04
ru	6.1	1	2500	151.63	1	386	20.12	1	45	1.85
de	2.67	1	2356	57.12	1	335	7.28	1	13	1.22
pt	1.95	2	2482	82.89	1	390	11.89	1	20	1.36
ko	1.91	1	1487	53.41	0	156	8.13	1	11	1.12
UNK	1.59	1	2500	32.37	0	364	3.82	1	28	1.1
ar	1.38	1	1552	61.14	1	230	10.12	1	21	1.12
it	1.14	2	2124	62.78	1	329	8.95	1	17	1.29
th	0.96	2	2426	74.22	1	405	14.47	1	9	1.12
fr	0.81	1	1482	51.51	1	198	6.67	1	8	1.19

The above observations not only advance our understanding of user behavior but also inform the design of machine learning models tailored to predict such behaviors. The knowledge of these distributions can significantly improve the accuracy and efficiency of targeted recommendations, whether they pertain to content, connections, or even advertising strategies. Hence, the insights gleaned from our cumulative user statistics serve as invaluable foundational data for both academic and practical applications in the realm of social network analytics.

B. Cross-Platform User Identification

In the realm of cross-platform social media engagement, our analysis revealed compelling figures. Specifically, 55% of users included their Instagram handle within their profile or biography, whereas 41% did so for Twitter. Interestingly, for 29% of these users, their Instagram and Twitter identifiers are identical. Such data serves as a crucial avenue for augmenting user information, broadening our understanding of their social circles, interests, and followership patterns. This underpins the potential for cross-platform analytics to offer a more comprehensive and multi-faceted user profile.

C. Multilingual Biography Analysis

We extended our analysis to user biographies written in multiple languages, unearthing a host of intriguing observations. Table I showcases the statistical breakdown of biography text attributes, delineating how text length, word count, and sentence structures vary across different languages. For instance, the average character count for an English-language biography stands at 107, while the corresponding figures for Korean and Russian are 52 and 152, respectively. Additionally, approximately 1% of biographies included an email address, suggesting a small yet potentially significant channel for direct communication or networking.

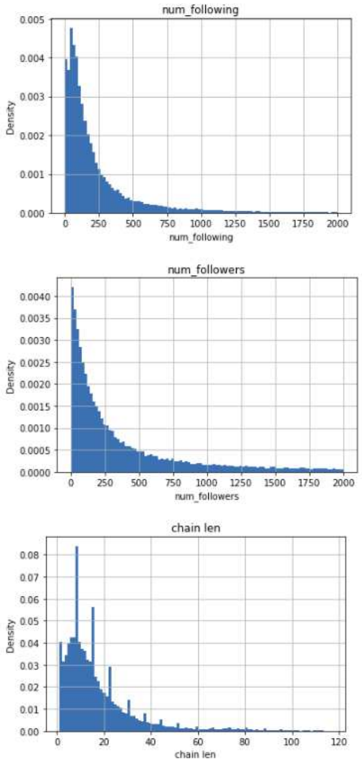


Fig. 4. Vital user data distributions (part 1)

The disparity in bio text attributes across languages offers valuable insights into the cultural and linguistic nuances that influence social media engagement. Moreover, understanding these variances can significantly enrich Natural Language Processing (NLP) models geared toward text analysis and summarization in multiple languages.

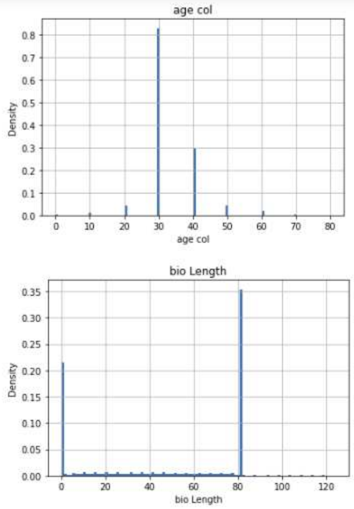


Fig. 5. Vital user data distributions (part 2)

The findings from our cross-platform identification and multilingual biography analysis serve as cornerstones for both academic discourse and practical applications, particularly in the optimization of targeted user engagement strategies and the development of more inclusive, multi-lingual machine learning models.

D. Graph and Image visualisations

1) Data Privacy Measures in Visualization

In our endeavor to delve deep into user relationships and social connections, we constructed comprehensive profile knowledge graphs. It's worth noting that stringent data privacy measures were observed: all personally identifiable information (PII), including full names and system identifiers, were meticulously stripped from our visualizations to ensure user anonymity.

2) Complexity and Interconnectedness in User Networks

Fig. 6 and Fig. 7 respectively present illustrative visualizations capturing a singular user's following network and the exponential complexity of connections when we expand to the entirety of his/her following network. The intricacies in these graphs serve as fertile grounds for identifying tightly-knit communities, as well as notable influencers within the network.

3) Cross-Platform Influencer Identification

By isolating the individuals who command the most significant following, we can extend our research to other social media platforms. This cross-platform approach can unearth a more nuanced understanding of an influencer's digital footprint and audience engagement metrics.

4) Graph Traversal and Storage Technologies

For graph traversal operations, we utilized Apache TinkerPop's Gremlin, a powerful and flexible graph traversal language [18]. This allows us to conduct complex analysis on our constructed graphs, such as identifying influencers and various special-interest groups.

5) Cloud-based Scalability

In terms of data storage and real-time query execution, we leveraged Amazon Web Services' Neptune, a fully-managed graph database service. This cloud-based solution provides robust scalability and is capable of handling thousands of Gremlin and SPARQL queries per second, thereby providing the computational power and speed necessary for our data-intensive tasks.

Our graph and image visualizations not only offer a detailed glimpse into the social dynamics at play but also provide a foundation for future research, particularly in the identification of influencers and community structures within social media networks. The use of advanced graph databases and traversal languages significantly contributes to the efficiency and scalability of our analytics architecture.

We were able to get detailed information about users, who they follow, obtain invitation chains and build profile knowledge graphs. To protect personal information, we have removed full names and other PII including system ids from our visualizations.

6) Enhanced Image Data for User Tracking

User-provided photos can serve as rich data sources, often embedding information such as geolocation coordinates and time stamps. Utilizing a limited number of these publicly available photos from social networks, we successfully mapped out user locations on the map (Fig. 8). It is critical to acknowledge that many individuals may prefer to keep such spatial and temporal data private.

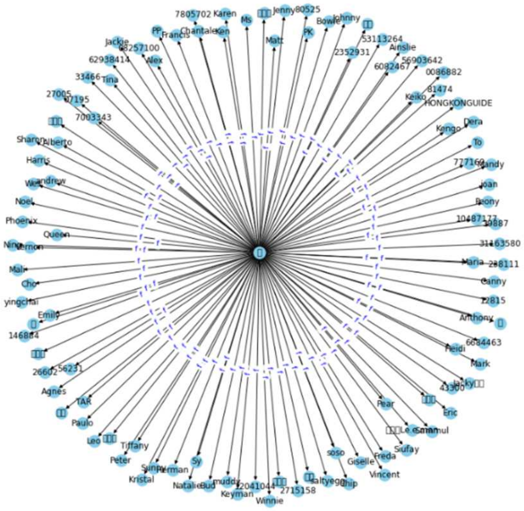


Fig. 6. Profile following user

When this geolocation information is combined with facial recognition technologies, the power to identify individuals—alongside details of their travel destinations and timelines—becomes possible. Thus, this raises essential ethical questions about user privacy and data confidentiality.

7) Graph-based Visualization of User Invitation Chains

Additionally, we have visualized the networks of invited profiles by traversing back to the "core" or root user who initiated the chain of invitations. Fig. 10 provides a visual representation of this invitation chain. In this graphical representation, each edge indicates the "distance," or relational length, to the originating profile.

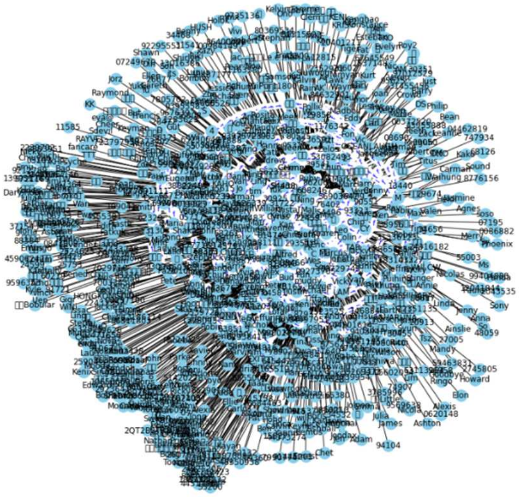


Fig. 7. Profile following for multiple users

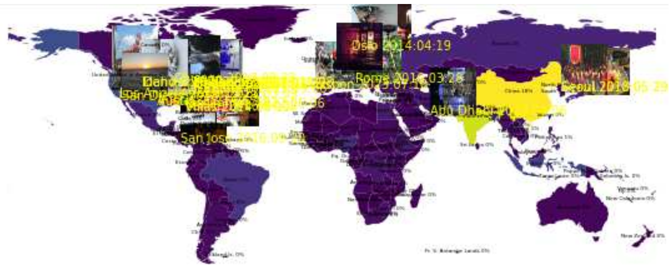


Fig. 8. Geolocation and timestamp of multiple photos for a user

The capability to map out invitation chains and to identify core users offers another layer of complexity to our understanding of social network dynamics. However, it equally brings up ethical considerations about user consent and data use.

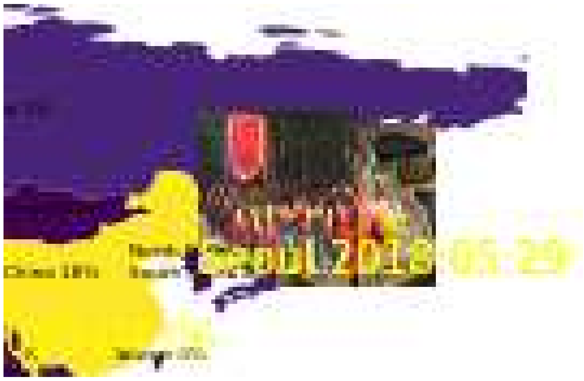


Fig. 9. Zoomed image showing city, date or any other photo metadata

Fig. 9 shows the zoomed version of the map containing metadata details from the photo. The picture displays city and date below the image. Potentially, we can display any other metadata including time, coordinates, or altitude.

In the culmination of our study, we successfully employed Natural Language Processing (NLP) techniques to visualize user profiles in a meaningful way. Fig. 12 demonstrates that accomplishment. Remarkably, the application of graph-based visualizations allows us to succinctly summarize and display the biographies of user profiles. By leveraging these visualizations, we can extract key topics of interest and activity patterns, thereby gaining nuanced insights into user behavior and preferences.

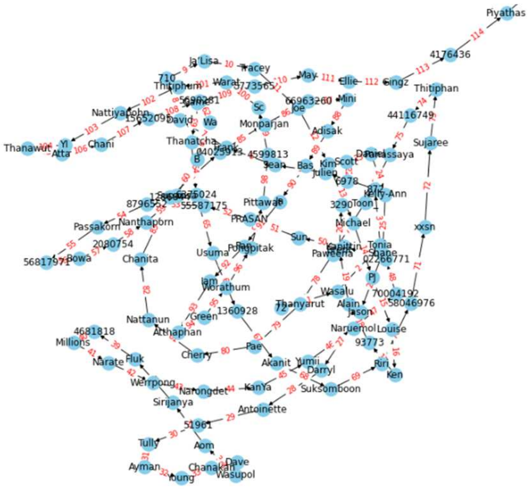


Fig. 10. User invited chain (over 100 links long)

8) Bio Summarisation and Topics detection using LLMs

The pace of Generative AI is amazing. Almost weekly, new models, tools or architectures are being released to tackle world-class problems. Almost daily, new use cases are being solved using diverse Large Language Models.

Our LLM approach (Fig. 11) was to design a serverless architecture that utilizes assorted generative models utilizing Lambda functions, AWS Bedrock (for inference and completion generation using Claude, Mixtral and Llama3 models), AWS Guardrails (for safe and secure AI, detecting and reporting toxic and sensitive topics), and Open AI API to call Chat GPT models.

Besides employing text-to-text models, we also leveraged multimodal image-to-text models, such as Anthropic's Claude 3 and Stable Diffusion, supported by AWS Bedrock. These models enabled us to accurately describe and classify images across various categories. For instance, we were able to provide detailed descriptions for images containing a single person, groups of people, objects, and natural landscapes. The combination of these models allowed for a comprehensive understanding of visual content, facilitating more nuanced and context-aware interpretations. This approach proved invaluable in applications ranging from automated content filtering to understanding users audience. The flexibility and scalability provided by AWS Lambda and Bedrock ensured that these models could be seamlessly integrated into diverse and efficient pipelines. Bedrock Guardrails helped us to implement safe and ethical analytics, prohibiting toxic and anti-Semitic content along with masking PII information from both prompt and completion.

We were able to successfully summarize all users' bios into English using several popular Generative AI models. In average, it took around two seconds to summarize text using OpenAI's GPT-4o model. Open source models performed really well. Falcon7b generated summaries within 100 milliseconds using powerful GPU, and Mistral7b within 10 seconds using llama_cpp on CPU only machine.

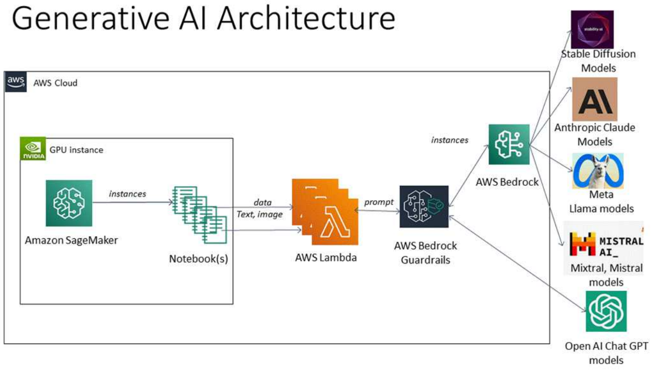


Fig. 11. Serverless Generative AI Architecture usilizing LLM models

We successfully clustered similar users by utilizing the K-Means algorithm. The input data consisted of ten topics in English, generated by Falcon7b LLM for each user's biography. We just created a vector of zeros and ones (containing detected subjects) for each profile and fed that data to the clustering algorithm. The efficiency and speed with which these methods handled texts in multiple languages, as well as their ability to perform complex NLP tasks such as translation, summarization, embedding, and topic detection within milliseconds, were notably impressive. Additionally, we have saved embedded biographies into vector database for quick search for topics, subjects and users in the RAG application [26].

IV. Conclusion

Despite the perception of privacy, the reality is that social networks are not as private as one might think. Unless a user has taken deliberate steps to maintain a confidential profile, a host of personal information—including friends, followers, photographs, and preferences—is readily accessible to external parties. To uphold ethical considerations, we have intentionally abstained from providing specific details about platforms' APIs in this paper.

Cloud computing serves as an ideal ecosystem for the secure scalable storage, processing, and automation of data. Serverless architectures, exemplified by Lambda functions, are automatically triggered by specified cloud events, enabling real-time processing of new profiles. The scalability of our serverless application adjusts dynamically based on computational load, thereby optimizing cost. Notably, Lambda offers a cost-efficient pricing model, where we incur charges only for the actual function calls, at a rate of $0.20 per two million invocations.

Rapid advancements in Generative AI and Large Language Models (LLMs) have greatly enhanced the capabilities for quick text understanding, summarization, and clustering, significantly transforming how data scientists and researchers interact with and process data. These technologies now enable complex textual analyses to be completed in mere seconds, revolutionizing data-driven workflows and insights. LLMs have substantially elevated the efficiency and sophistication of text processing tasks.

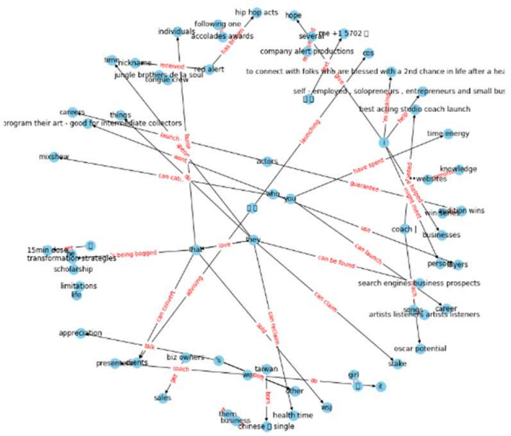


Fig. 12. User bio text visualization with graphs

Modern multimodal LLMs have the capability to describe user images, unlocking limitless possibilities for innovative analytical applications across a wide range of domains. These models enable precise classification of photos, allowing for the identification and filtering of human faces, objects, and other elements within the images. This advanced capability not only enhances image analysis and categorization but also supports applications such as privacy protection, content moderation, and automated tagging. By accurately discerning and isolating specific components in images, these models drive more efficient data processing and provide deeper insights into visual content.

While social networks do offer APIs for third-party integration to boost traffic and profitability, the platforms must exercise due diligence in safeguarding what information is disclosed through their APIs.

User awareness is equally crucial. Our analysis found that many individuals freely publish their email addresses and social media handles within their profiles or bios. Such practices make it exceedingly easy for malicious actors to acquire and misuse this personal data. This vulnerability is not limited to any particular platform; even paid premium services can potentially expose your email, phone number, or physical address. Users should exercise extreme caution in disclosing any Personally Identifiable Information (PII) unless absolutely necessary.

References

[1] Scott Fulton III, “To be a microservice: How smaller parts of bigger applications could remake IT”, https://www.zdnet.com/article/to-be-a-microservice-how-smaller-parts-of-bigger-applications-could-remake-it/

[2] Kaplunovich, Alexander, “*Real-Time Automatic Hyperparameter Tuning for Deep Learning in Serverless Cloud*”, University of Maryland, Baltimore County, 2020.

[3] Arvind Narayanan and Vitaly Shmatikov, “How To Break Anonymity of the Netflix Prize Dataset (2006)”, CoRR, http://arxiv.org/abs/cs/0610105, 2006.

[4] Danilak, M, "langdetect: Language detection library ported from Google’s language detection", *See https://pypi. python. org/pypi/langdetect/(accessed 19 January 2015)* (2014).

[5] Lui, Marco, and Timothy Baldwin, "langid. py: An off-the-shelf language identification tool", *Proceedings of the ACL 2012 system demonstrations*. 2012.

[6] M. DeWilde, Burton, "Textacy: NLP, before and after spaCy." (2020).

[7] Yamada, Ikuya, et al, "LUKE: deep contextualized entity representations with entity-aware self-attention", arXiv preprint arXiv:2010.01057 (2020).

[8] Vision Transformer (ViT) documentation, Huggingface, https://huggingface.co/docs/transformers/model_doc/vit

[9] Zhou, Daquan, et al, "Deepvit: Towards deeper vision transformer", arXiv preprint arXiv:2103.11886 (2021).

[10] Peter Sbarski, “Serverless Architectures on AWS”, Manning 2017

[11] Juarto, Budi, and Abba Suganda Girsang, "Neural Collaborative with Sentence BERT for News Recommender System", JOIV: International Journal on Informatics Visualization 5.4 (2021): 448-455.

[12] Reimers, Nils, and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", arXiv preprint arXiv:1908.10084 (2019).

[13] Schroff, Florian, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering", Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[14] face-compare documentation, https://pypi.org/project/face-compare/

[15] Huang, Kezhao, et al, "Understanding and bridging the gaps in current GNN performance optimizations", Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. 2021.

[16] Bronstein, Michael M., Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst, "Geometric deep learning: going beyond euclidean data", IEEE Signal Processing Magazine 34, no. 4 (2017): 18-42.

[17] Hagberg, Aric, Pieter Swart, and Daniel S Chult, “Exploring network structure, dynamics, and function using NetworkX”, No. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[18] Gremlin documentation, https://tinkerpop.apache.org/gremlin.html

[19] Kaplunovich, Alex, "ToLambda--Automatic Path to Serverless Architectures", 2019 IEEE/ACM 3rd International Workshop on Refactoring (IWoR). IEEE, 2019.

[20] Jayaram, B., and C. Jayakumar. "A survey on security and privacy in social networks." Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2021. Singapore: Springer Singapore, 2022. 807-822.

[21] Kumar, C., et al.: Online social network security: a comparative review using machine learning and deep learning. Neural Process. Lett. (2021)

[22] Gao, Yuan, et al. "IEEE access special section: privacy preservation for large-scale user data in social networks." *IEEE Access* 10 (2022): 4374-4379.

[23] Fire, Michael, Roy Goldschmidt, and Yuval Elovici. "Online social networks: threats and solutions." IEEE Communications Surveys & Tutorials 16.4 (2014): 2019-2036.

[24] Felt, Adrienne, and David Evans. "Privacy protection for social networking APIs." *2008 Web 2.0 Security and Privacy (W2SP’08)* (2008).

[25] Llama.cpp python library web documentation, https://python.langchain.com/docs/integrations/llms/llamacpp

[26] Chacko, Neha, and Viju Chacko. "Paradigm shift presented by Large Language Models (LLM) in Deep Learning." *ADVANCES IN EMERGING COMPUTING TECHNOLOGIES* (2023): 40.

[27] Kaplunovich, Alex, and Sophia Kaplunovich. "Consolidating user data from social networks using Machine Learning and Serverless Cloud." In 2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS), pp. 230-236. IEEE, 2023.