# A Risk-Based AML Framework: Finding Associates Through Ultimate Beneficial Owners

Sasan Jafarnejad, François Robinet, Raphaël Frank
Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg
29 Avenue J.F Kennedy, L-1855 Luxembourg
{sasan.jafarnejad, francois.robinet, raphael.frank}@uni.lu

***Abstract*—The ever increasing regulatory requirements for Anti-Money Laundering (AML) compliance presents significant challenges for financial institutions and small businesses worldwide. Efficiently navigating these requirements is crucial not only for legal adherence but also for safeguarding the integrity of the global financial system. In response to this challenge, we develop a framework that uses advanced algorithms to improve identification and risk assessment processes within Know Your Customer (KYC) procedures. Using a technique to measure graph-based node similarities, our approach enhances the detection of Politically Exposed Persons (PEPs) and their known associates, facilitating a more nuanced and comprehensive analysis than traditional methods allow. We study the dataset of Ultimate Beneficial Owner (UBO) registry in Luxembourg and translate our findings into two risk indicators: involvement with underage shareholders, and number of companies at the address. We integrate these two indicators, as well as several other components of AML compliance, including country risk indices, beneficial ownership structures, and adverse media exposure, into a single coherent risk metric. The framework is designed to be both modular, supporting various degrees of regulatory scrutiny, and scalable, suitable for evolving regulatory landscapes. This risk metric can be used to determine whether Enhanced Due Diligence (EDD) is required by European AML directives. The end result is a more robust defense against financial crimes and improved AML processes within the EU and beyond.**

***Index Terms*—SimRank, Anti-Money Laundering, Risk Assessment, Know Your Customer (KYC)**

## I. Introduction

The United Nations Office on Drugs and Crime reports that money laundering represents 2-5% of the world's GDP, fueling crime networks and funding terrorism [1]. Although this is a global issue, it is estimated that in Europe, only about 1. 1% of these funds are ever captured. This underscores the urgency and significance of robust AML and Combating Financing of Terrorism (CFT) measures [2]. Directive (EU) 2018/843 of 30 May 2018 also known as the Fifth Anti-Money Laundering Directive (AMLD) (5AMLD), enacted by the European Union (EU) in 2018, represents a significant advance in the region's efforts to combat money laundering and terrorism financing. Building on its predecessor, 5AMLD introduces new measures to enhance transparency, expand the scope of regulated entities, and strengthen due diligence requirements. By implementing rigorous AML/CFT protocols, European authorities not only safeguard the continent's economic stability but also contribute to global security by thwarting the financial lifelines that underpin terrorist organizations.

5AMLD notably expands the regulatory perimeter to include virtual currencies and prepaid payment methods. It also mandates the creation of publicly accessible national UBOs repositories, thereby bolstering the commitment to financial transparency and accountability of the EU. These directives oblige not only Financial Institutions (FIs) to perform costly and time-consuming compliance checks, but also require small businesses such as accountants, art dealers, or real estate agencies to follow the same obligations. These compliance checks are known as Customer Due Diligence (CDD) and sometimes generally referred to as KYC. A follow-up directive further expanded the regulatory scope so that *aiding and abetting* will now be punishable as criminal offenses. Therefore, the *enablers* will be equally guilty. This poses additional risks to Small and Medium Businesses (SMBs) that do not have the capacity to allocate to KYC operations.

Additionally, 5AMLD emphasises importance of UBOs and expands the definition of what would be considered a PEP. These changes are supported by research on the importance of transparency of beneficial ownership data for the identification and prevention of illicit financial flows. As an example, in his comprehensive analysis, Sharman examines the global standard on transparency of beneficial ownership and its implementation across jurisdictions [3]. Other studies have also identified that the complexity and opacity of corporate structures are significant barriers in the detection and prosecution of money laundering activities. A report by the World Bank Group delves into the misuse of legal entities and the challenges facing AML frameworks [4]. Their work is pivotal in understanding the methods used by launderers to obscure true ownership, thereby informing the development of more targeted regulatory tools.

AMLDs foresee a more comprehensive form of CDD, called EDD, in special cases including high-risk individuals. Specifically, 5AMLD broadens the scope of EDD to include not only foreign PEPs but also domestic PEPs. That means transactions involving PEPs require EDD and ongoing monitoring; the directive also extends EDD to family members and known associates of PEPs. This is particularly challenging since there are no harmonized databases for PEPs nor common definitions for terms such as *known associates* between countries.

Furthermore, FIs are guided to adopt a risk-based approach when dealing with PEPs, this involves comprehensive risk assessments, which take into account variables such as country of origin, the nature of public function including industry, and related corruption indexes.

Identifying and assessing the risk of money laundering through analysis of UBO networks is an emerging approach in AML compliance. Traditional rule-based systems often struggle to detect complex money laundering schemes, leading to high false positive rates and operational inefficiencies [5]. Recent studies have explored the use of Machine Learning (ML) and graph analysis techniques to enhance AML detection in UBO networks. A proposed method constructs a hierarchical risk control knowledge graph (HRCKG) from simulated AML data, extracting transaction features to automatically generate risk control rules [6]. The HRCKG enables rule-based and graph-based reasoning to assess account-level money laundering risk and identify suspicious groups. Network analysis has also been applied to real-world banking data to detect anomalies indicative of money laundering. One approach designs new centrality features based on ego networks and random walks to capture circular transaction flows [7]. These features, combined with an unsupervised anomaly detection algorithm, demonstrate strong performance on real and synthetic data. While transaction-based ML approaches have been studied for AML, risk assessment based solely on UBO data requires different techniques and faces unique challenges regarding data quality and regulatory compliance. Some work has focused on modeling UBO networks to identify high-risk relationships. A relational model was proposed using social network analysis (SNA) techniques to discover connections between suspicious customers in an AML context [8]. This model provides a framework for uncovering hidden risk patterns in complex ownership structures.

In this work, our aim is to address some of the challenges introduced by the ever-expanding compliance requirements. More specifically, our focus is on newly introduced requirements regarding PEPs, and UBOs. We accomplish this by introducing an extensible risk-based framework for KYC. Our four core contributions are as follows:

1) Proposing a novel approach to identifying *known associates*, based on SimRank [9].
2) Releasing a Graphics Processing Unit (GPU) accelerated implementation of SimRank[1].
3) Introducing a risk metric drawing on the guidelines presented in the AML directives, as well as our observations from comprehensive analysis of UBO registry of Luxembourg.
4) Releasing anonymized dataset of UBO registry of Luxembourg, to the research community[2].

The rest of the paper is structured as follows, in Section II we present the UBO dataset, the anonymization process and present an exploratory data analysis. In Section III we introduce our proposed KYC approach, including details on the calculation of risk metric and the role of SimRank. Next, we evaluate the proposed method in Section IV and present the results. In Section V we discuss our findings. We conclude in Section VI

[1]Source code available at https://github.com/sasan-j/simrank-cuda.git

[2]https://www.kaggle.com/datasets/sasanj/ultimate-beneficial-owners-companies-investments

## II. Dataset and Insights

The 5AMLD obliged EU member states to establish a register of UBOs available to the general public and other member states. Luxembourg followed suit and opened up its *Registre Des Bénéficiaires Effectifs (RBE)* in 2019. The data used in this work is an extract of the *RBE* which was obtained in 2021, and we refer to this as the dataset in the rest of the paper. However, a ruling by the European Court of Justice [10] recognized that public access to these registers infringes on the rights granted to individuals through General Data Protection Regulation (GDPR), as a consequence the RBE is no longer accessible by the public but limited to professionals such as those working at FIs.

The original, non-anonymized, dataset contains the following information:

1) For each company: name, company identifier, list of administrators and shares of UBOs. Using the company ID, we could expand the dataset by querying the public company register and add the company address, legal form, and, *Nomenclature statistique des Activités économiques dans la Communauté Européenne (NACE)* which identifies the sector of activity.
2) For each of the individuals (administrator or UBO): name, nationalities, birth date, birthplace.

The dataset consists of 93919 companies and 71459 UBOs. Note that this is a loose use of the term, since in cases such as trusts only people marked as *Manager* appear in the dataset. In Luxembourg and in Europe, each company has a NACE code which determines the sector in which the company operates. This allows us to categorize the companies and their respective industries.

### A. Anonymization Process

The dataset used in this work was anonymized as a result of the above-mentioned ruling, leading to the closure of the register. We take several steps to anonymize the dataset. First, we replace the name of all people and companies with a unique random identifier. This data would be useful for correlating with external datasets, but it would also allow the individuals to be re-identified. We do not maintain any cross-reference or lookup table that would allow someone to reverse-engineer these IDs back to the original identifier. It is impossible to link the ID to the name after the fact. We also eliminated the company ID. Although companies are not covered by the GDPR, not anonymizing them would open the possibility of reidentification using external data sources (Linkage Attack).

Company addresses are replaced with unique random IDs. This allows us to link (anonymized) companies with a given (anonymized) address but not to link the address back to

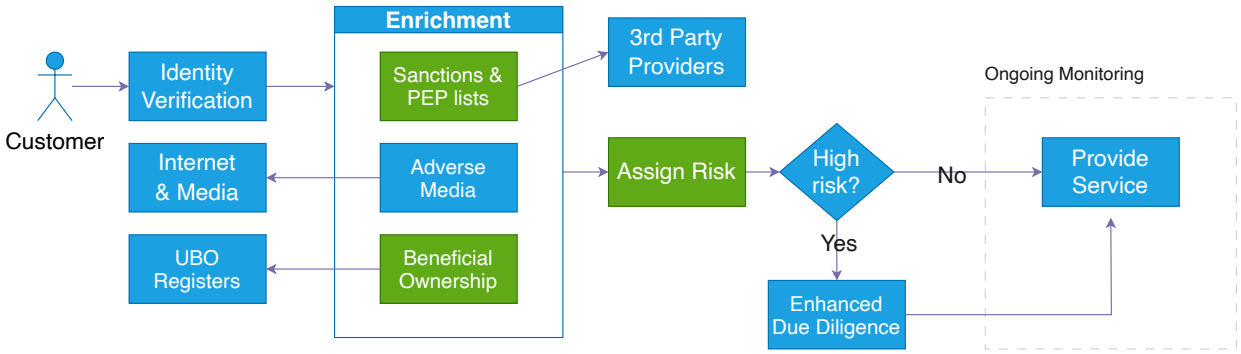


Fig. 1. The proposed KYC Process. The work presented here focuses on the green boxes. At the enrichment stage, we use SimRank to uncover known associates by bridging, PEP lists, and beneficial ownership information. We also propose a composite risk score as a flexible and novel risk assessment metric.

the company. Exact birth dates are replaced with the year of birth, in order to be able to study the effect of age on the data without being able to link the birth date back to a specific person. The combination of these techniques ensures that the representatives and the UBOs remain unidentified and protected.

### B. Exploratory Data Analysis

This UBO dataset is particularly rich in information and could lead to many different studies. In this work, we focus our analysis on two risk factors that are particularly good predictors in the context of AML/CFT.

*1) Companies with Underage Shareholders:* In Luxembourg, individuals under the age of 18 are considered minors. Nevertheless, our dataset lists 122 minors as ultimate beneficial owners. There are 104 companies that have at least one minor shareholder, and 21 companies with only minors listed as their UBOs. Though this is not illegal, it is considered as an indicator for concealment of beneficial ownership by the Financial Action Task Force (FATF). Initially we suspected that these companies may mostly be patrimonial companies (e.g., SOPARFIs and SPFs), but later found that 89% of them are Public Limited Companies (SAs) (26%) or Limited Liability Companies (SARLs) (62%). Note that this fraction of SAs and SARLs is the same in the rest of the dataset. Presence of companies with minor shareholders has sparked the interest of other researchers as well, which proves its significance [11].

*2) Registered Address:* We observed that it is not uncommon for hundreds of companies to be registered at the same address. This maybe common in two general cases: in cities with large skyscrapers or in countries where letter box companies are allowed. However, this is not case in Luxembourg. There are 17 addresses with over 500 companies registered (3 with over 1000 companies), spread over four localities namely, Luxembourg City, Senningerberg, Bertrange, and Leudelange.

Looking at the companies housed in the top three address which constitute 4501 companies, we notice that we have ownership information for just over 19% of them, while the same statistic for all other companies is 69%. Note that in Luxembourg letter box companies are not permitted, except for patrimonial companies (SOPARFIs and SPFs) and dormant companies. The collected data shows that a vast majority of these 4501 companies (90%) are using either the SARL or SA legal form. Looking at the prior public leaks, we can find 4 companies that are mentioned in the *Paradise papers* [12] and *Panama papers* [13], are registered in 2 of these 3 top addresses. This finding does not allow to conclude that companies registered at these addresses are involved in illegal activity, but shows that the number of companies registered at a given address can be used as one useful indicator in a risk-based assessment.

## III. Methodology

We start this section by describing our proposed KYC process depicted in Figure 1. Our aim is to introduce mechanisms that can be plugged into existing KYC systems or used as building blocks for an entirely new KYC approach. With respect to identifying PEPs, especially the notion of *close associates*, we propose an approach to close the gap between automated KYC systems and the guidelines set forth by AMLD and FATF. We apply SimRank [9] on the bipartite graph of companies and their UBOs, in order to identify related entities. Additionally, we propose the use of risk metrics that can be expanded and tuned to the use-case and risk appetite of each FI, as a way to model the risk-based approach to AML/CFT recommended by AMLD and FATF guidelines.

First, identification documents are provided for a legal or natural entity (i.e., a company or a person). Although the methodology can be applied to either legal or natural entities for simplicity in the rest of this work we focus only on natural persons. We assume establishing the identity of the customer is a solved problem, this is evident by myriad of companies providing such services (e.g. Veriff[3], Onfido[4], etc.) Our focus is on the *enrichment* and *risk assignment* steps.

### A. Enrichment

In Figure 1, we highlighted 3 parallel steps for enrichment. Although not exhaustive, this is what most solutions cover.

[3]https://veriff.com
[4]https://onfido.com

The goal is to enrich the identification document with supplementary information, to use in the next stage to assign risk to the entity and decide whether we require to perform EDD.

Here most solutions generally consider cross referencing the name of the entity with various international datasets, such as sanctions lists provided by organizations such as United Nations, EU, Office of Foreign Assets Control (U.S.), Her Majesty's Treasury (U.K.), as well as PEPs lists. Additionally, news sources are scanned for negative or alarming articles related to the entity. The aim is to identify whether the individual is a PEP or high-risk entity.

Our proposal is more comprehensive, since we also take into account entities related to the individual in question. Recall that 5AMLD obliges member states to setup UBOs registers and make them accessible. As presented before for this study we only have access to the Luxembourg register, but we build upon this dataset knowing it could be expanded to countries across the EU.

In order to uncover complex links between individuals investing in similar companies, and companies sharing similar shareholders, we use the SimRank algorithm [9]. SimRank is designed to measure the similarity of nodes in a graph based on their structural context. The main idea is that two nodes are considered similar if they are related to similar nodes. The original SimRank formula for two nodes $i$ and $j$ is computed recursively using the following equation:

$$S(i,j) = \frac{C}{|N(i)| \times |N(j)|} \sum_{a \in N(i)} \sum_{b \in N(j)} S(a,b)$$

Where, $S(i,j)$ is the similarity between nodes $i$ and $j$. $N(i)$ is the set of in-neighbors for node $i$. $C$ is a constant factor to decay the similarity value. The sums iterate over all in-neighbors $a$ of $i$ and $b$ of $j$.

Weighted SimRank enhances the original formula by considering the importance of neighboring nodes and the structural information of the graph. The rationale is that not all neighbors contribute equally to the similarity score. In the context of UBOs, a minority shareholder is not equally as important as a majority shareholder.

Weighted SimRank introduces a weight for each neighboring node pair $(a,b)$ which reflects their importance. The formula is adjusted to the following:

$$S(i,j) = \frac{C}{|N(i)| \times |N(j)|} \sum_{a \in N(i)} \sum_{b \in N(j)} W(a,b) \times S(a,b)$$

Where $W(a,b)$ is the weight of the neighboring node pair $(a,b)$.

Applied to our problem, SimRank is able to uncover similarities between investors even when they do not invest in the exact same companies, as depicted on Figure 2. Because there exists two types of vertices (nodes), namely UBOs or investors and companies, we consider a specific variation of SimRank for bipartite graphs. We use the term investor loosely here, since in the dataset we have a large number of management directors that are listed as UBO. Formally, let $G = (V, E)$ be the bipartite investment graph, where $V$ is the set of vertices and $E$ is the set of edges. The vertex set $V$ is divided into two disjoint subsets: $V_I$ and $V_C$. $V_I$ represents the set of investors. Each vertex $i \in V_I$ corresponds to a unique investor. $V_C$ represents the set of companies. Each vertex $c \in V_C$ corresponds to a unique company. Thus, $V = V_I \cup V_C$ and $V_I \cap V_C = \emptyset$. The edge set $E$ consists of edges that connect the vertices between these two sets. An edge $(i,c) \in E$ exists if and only if investor $i$ has made an investment in company $c$. Furthermore, we introduce a weight function $W : E \to \mathbb{R}$, which represents the share ratio of the investor in the company (1.0 indicating 100% ownership). However, in the case that the share information is not available, we assume equal ownership among all managers.

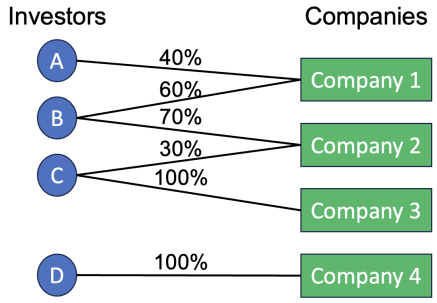


Fig. 2. Example Investment Graph. The percentages on the edges indicate the ownership. If we apply SimRank on this graph, and use the ownership as weights, we observe that, $A, B, C$ have no similarity to $D$. On the other hand $A$, and $C$ will have a non-zero similarity, since both have $B$ as a co-investor.

$$S(i,i') = \frac{C_1}{|N(i)||N(i')|} \sum_{c \in N(i)} \sum_{c' \in N(i')} W(i,c) W(i',c') S(c,c') \tag{1a}$$

$$S(c,c') = \frac{C_2}{|N(c)||N(c')|} \sum_{i \in N(c)} \sum_{i' \in N(c')} W(c,i) W(c',i') S(i,i') \tag{1b}$$

Therefore, two companies are similar if they have similar investors. And two investors are similar if they invest in similar companies. This recursive formulation is solved through an iterative approach, with an initial condition that makes the nodes only similar to themselves. Considering that we have around 52k investors and 75k companies, due to nonlinear time and space complexities (see the SimRank paper [9]), running the algorithm for such a large graph on a desktop computer is not feasible. Approaches using dense matrices yield results in minutes but require large amounts of memory, while using sparse matrices resolves the memory issue but takes in order of hours of runtime. To address this, we implemented a parallelized version of SimRank, accelerated by GPU using CUDA. This allowed us to run the computations in seconds; however, it was still challenging to run the program for the entire graph on a single NVIDIA A100 GPU equipped with 32GB of memory. However, the remedy was quite simple. Note that in the real world there is no single investor with investments in thousands of companies and no company that has thousands of investors. We observe that the graph of investors and companies is composed of 35750 disconnected components, of which 19385

are components consisting of a single company with a single investor, while the largest component contains 45167 investors and companies. By definition, nodes belonging to different connected components have a similarity of 0. Therefore, by splitting the graph into several disconnected graphs we can run SimRank much faster and combine the results back in an efficient data structure such as hash-map or a sparse matrix.

What we end up with is a way to find similar people among the UBOs, and this similarity is quantified with a value in the range $0 \leq S < 1$, where 0 indicates no similarity and 1 is the maximum similarity.

The next challenge is to obtain a comprehensive PEP and sanctions list. Although real-world applications use commercial aggregators, the anonymity of our dataset prevents the use of such data. For brevity, we use *PEP list* to refer to a complete collection of high-risk lists such as sanctions, PEPs, and warrants. Conventional approaches simply cross-refer customers in these lists, but we aim to address the challenge of identifying close associates.

When onboarding a new customer, we check for a hit on the PEP list. If found, we set the $PEP$ flag to *true*. We also retrieve the list of people who have similarity of above a chosen *similarity threshold* to the customer. We call this list *known associates* represented by $N_A$. We also look up the PEP list for each known associate and flag them if we get a hit.

Now that we enriched the customer data using UBO register data and PEP lists and established the known associates, we proceed to the next step to assign risk to the individual.

### B. Risk Assignment

The construction of a risk metric for the KYC processes requires a methodical approach that quantifies various risk factors and integrates them into a cohesive metric. This metric aims to encapsulate the multifaceted nature of AML/CFT risks. The risk metric, denoted as $R$, is defined by the following equation:

$$R = \alpha \cdot \text{CRI} + \beta \cdot \text{AOI} + \gamma \cdot \text{PSLI} + \\ + \delta \cdot \text{MII} + \epsilon \cdot \text{AMI} + \zeta \cdot \text{R}_{\text{associates}}$$

Where, CRI stands for Country Risk Index, a score based on risk from country of citizenship and country of birth. AOI is the Address Overlap Index, indicating the number of companies at the registered address, normalized against a threshold. PSLI is the PEP and Sanctions List Index, reflecting direct association with any individual or entity on the PEP and sanctions lists. MII stands for Minor Involvement Index, which captures the presence of minors in companies with which the person is involved. AMI stands for Adverse Media Index, capturing the presence in negative news reports or investigative databases. $\text{R}_{\text{associates}}$ is the risk from associates identified using SimRank.

To calculate the risk metric ($R$) we make some assumptions and obtain concrete values for each component.

We calculate CRI based on the on FATF country evaluations [14] and Corruption Perceptions Index (CPI) [15]. These parameters are extracted from both the citizenship countries and the place of birth. For the vast majority, country of birth is among the countries of citizenship; however, for the few that this is not the case, it should be considered in the risk score. This is particularly important if one of the nationalities or country of birth is a high-risk jurisdiction.

$$\text{CRI} = \max_{c \in (C_{\text{citizenship}} \cup C_{\text{birthplace}})} (R_{\text{FATF}}(c) + R_{\text{CPI}}(c))$$

Where $R_{\text{FATF}}(c)$ is defined as:

$$R_{\text{FATF}}(c) = \begin{cases} 1 & \text{if } c \text{ in FATF black list,} \\ 0.5 & \text{if } c \text{ in FATF gray list} \end{cases}$$

And $R_{\text{FATF}}(c)$ as:

$$R_{\text{CPI}}(c) = 1 - \frac{CPI(c)}{100}$$

$CPI(c)$ is the country CPI score according to [15]. The choice of constants are somewhat arbitrary and can be adjusted by domain experts; however, our goal is that cases which should require EDD according to AMLD or FATF guidelines will score 1 or higher. Therefore, we design with a threshold of 1 in mind to trigger the EDD. This formulation results in 1676 (2.3%) of the population scoring RCI $\geq 1$, while the majority (87%) score less than 0.5.

To calculate AOI we use the following formula:

$$AOI = \min\left(\frac{\max_{i \in \text{Investments}} (\text{\# companies\_at\_address}_i)}{K}, 1\right)$$

Given our dataset, we chose the normalization factor of $K = 1000$. We chose 1000 because our dataset contains only 3 addresses at which more than 1000 companies are registered. This can be adjusted by domain experts to the appropriate value for any dataset. We see that with this definition 23565 (33%) individuals will get the maximum score of 1. This seems to be quite odd, since the entire dataset contains just about 70k individuals. Recall that there are 3 addresses with more than a thousand registered companies. Those companies also have a larger number of UBOs, 2.62 in average versus 2.04 for all other companies.

We assign MII to 1 if the individual is UBO in any company with minor UBOs, and 0 otherwise. Only 395 (0.5%) UBOs get flagged with this index.

In our experiments, we exclude AMI, since it is out of the scope of our work.

We define the function $f(a)$ that represents the risk of an associate. Where $a$ is an associate within the set of known associates $N_A$. $N_A$ is computed as the set of all UBOs that have a SimRank similarity score above the similarity threshold with the individual for which risk is evaluated.

$$\text{R}_{\text{associates}} = \sum_{a \in N_A} f(a)$$

Where $f(a)$ is defined as:

$$f(a) = \begin{cases} 1 & \text{if associate } a \text{ is flagged,} \\ 0 & \text{otherwise.} \end{cases}$$

The function $f(a)$ acts as an indicator function that maps the condition of being flagged to the binary outcomes required for

TABLE I
SUMMARY OF EXPERIMENTS AND RESULTS

| # PEPs (% Population) | SimRank Variant | $\#R_{\text{associates}} \geq 1$ Mean $\pm \sigma$ | $\# R \geq 1$ Mean $\pm \sigma$ | $\# R \geq 2$ Mean $\pm \sigma$ |
|---|---|---|---|---|
| 25 (0.03%) | Unweighted | 151.9 $\pm$ 65.2 | 2042.9 $\pm$ 63.3 | 10.7 $\pm$ 2.2 |
| 25 (0.03%) | Weighted | 14.3 $\pm$ 4.5 | 1919.9 $\pm$ 4.1 | 7.3 $\pm$ 0.6 |
| 50 (0.07%) | Unweighted | 294.3 $\pm$ 105.1 | 2179.8 $\pm$ 100.3 | 18.9 $\pm$ 24.2 |
| 50 (0.07%) | Weighted | 31.4 $\pm$ 7.0 | 1926.6 $\pm$ 6.9 | 7.6 $\pm$ 0.8 |
| 100 (0.14%) | Unweighted | 590.6 $\pm$ 124.1 | 2465.8 $\pm$ 119.7 | 39.0 $\pm$ 38.4 |
| 100 (0.14%) | Weighted | 61.6 $\pm$ 9.8 | 1955.9 $\pm$ 9.5 | 8.3 $\pm$ 1.4 |
| 250 (0.35%) | Unweighted | 1445.9 $\pm$ 170.6 | 3290.5 $\pm$ 163.3 | 115.2 $\pm$ 60.43 |
| 250 (0.35%) | Weighted | 155.0 $\pm$ 15.8 | 2046.5 $\pm$ 15.9 | 11.9 $\pm$ 2.9 |

the summation. The sum $R_{\text{associates}}$ will give us the total number of flagged associates within the set $N_A$. Note that $\text{R}_{\text{associates}}$ can easily extended to cover other related entities that were obtained through other means. The weights ($\alpha, \beta, \gamma, \delta, \epsilon$ and $\zeta$) are assigned to each index reflecting their relative importance in the risk assessment process. They are determined based on regulatory guidance, expert input, and historical data analysis. For instance, $\alpha$ could be higher for countries with known deficiencies in AML controls, while $\beta$ might be increased if $AOI$ is a strong predictor of risk based on the institution's experience.

In the simplest case, when the risk score $R$ is calculated, it can be compared against a predefined risk threshold $R_{threshold}$ (e.g. 1). Those with risk score $R > R_{threshold}$ will be flagged for EDD. The threshold can be selected by analyzing historical cases and tuned to balance between identifying potential risks and operational efficiency.

The risk metric is designed to be dynamic, allowing periodic recalibration of weights and thresholds based on ongoing monitoring and the emergence of new risks. ML techniques can be applied to refine the predictive accuracy of $R$, using training datasets from historical KYC cases.

This risk metric formula embodies a robust and flexible approach to risk assessment in the KYC process, allowing nuanced capture of risk factors pertinent to AML/CFT.

## IV. EVALUATION AND RESULTS

We run experiments to evaluate the proposed concept; however, due to anonymization of the dataset and GDPR concerns we are unable to cross-correlate the UBOs with real PEP lists. Instead, we simulate this by marking random subsets of UBOs in our dataset as if they appeared in such lists.

For our experiments, we consider 4 scenarios, randomly flagging 25, 50, 100, and 250 of the individuals in the dataset as PEP. We repeat the experiment for each scenario with 30 random samples and report the average and standard deviation.

Additionally, we run all 4 scenarios, with the weighted and unweighted variants of SimRank, in order to study their respective impact on the results. The only difference between the two variants is that in the unweighted variant we remove the $W$ terms from Equations 1a and 1b, thereby only focusing on the existence of a UBO and ignoring the size of their ownership.

Table I contains the summary of the results of our experiments. We track the number of associates (# $R_{\text{associates}} \geq 1$) for each scenario as an indicator of the soundness of our approach to determine *known associates*. We observe that the increase in # of PEP only results in a linear increase of # $R_{\text{associates}} \geq 1$, which shows that our approach is scalable, avoiding an exponential growth which would be required to quickly perform EDD for the entire dataset.

For the risk score ($R$), we present the number of people scoring more than 1 and 2 in Table I. We can see that for the 4 different cases ranging from 25 to 250 PEP the number of individuals scoring above 1.0 ranges from 2042 to 3290 in the unweighted variant and from 1919 to 2046 in the weighted variant. However, the numbers are much smaller for the case of $R > 2$. Therefore, in practice, we can also consider using a laddered approach. For example, those with a score greater than 2 are sent directly for EDD while those with a score $1 \leq R < 2$ referred to be triaged or passed through extra but still automated checks.

Overall, we observe that the unweighted variant results in much larger number of individuals to be flagged for EDD ($R > 1$), this can be explained as the influence of weight terms (based on ownership ratios) which can dilute the similarity much faster, therefore far fewer individuals would be considered to be similar to each other. This can be alleviated by choosing a larger similarity threshold to reduce the number of known associates.

The framework provided by the Weighted SimRank is quite flexible and can be further expanded to include additional factors in the $W$ functions, such as industry and country of citizenship, to signal a higher degree of similarity.

## V. DISCUSSIONS AND FUTURE DIRECTIONS

In this section, we highlight some of the shortcomings of our work and propose ideas to address them in future research. Our study is comprehensive, yet it overlooks various elements of the KYC/AML process, which we will discuss now. First, the evaluation of *adverse media* coverage, represented by AMI in our risk metric, requires the complex retrieval and summarisation of the media, which may contain text, audio, and video. Second, *ongoing monitoring*, which includes both the monitoring of structured data (transactions) or PEP lists, as well as unstructured data, such as media (same as *adverse media*). Lastly, *behavioral profiling*, AML directives call upon FIs to build profiles for PEPs by determining their industry, identifying behavioral patterns, and to ensure extra vigilance when they deviate from their usual patterns. All of these share a similarity: they require the processing of various data sources with multiple modalities. To address this, tools such as multimodal Large Language Models, can be combined with methods such as Retrieval-Augmented Generation [16] to distill vast amounts of data into short summaries that will be reviewed by compliance officers.

We also note that the use of nominees and relatives by PEPs and high-risk individuals is a well-recognized means of money laundering [17], although we did not address this issue, we see it as an open area of research, looking at patterns such as birthplace, age, and name structure, perhaps augmented with media reports, or social media it may be possible to identify relatives or known associates with better recall, even though this approach may also have a high false positive rate, which needs to be addressed. As financial networks continue to evolve in complexity, the integration of sophisticated visualization tools into compliance workflows is becoming increasingly indispensable. The risk score and similarities found between companies and individuals can be improved with contextual information and visualized using interactive tools. This allows compliance officers to more effectively sift through data, discerning potential risks and threats with greater accuracy and speed. This not only improves the efficiency of AML/CFT measures, but also facilitates a more proactive approach in the identification of suspicious activity.

## VI. Conclusion

In this paper, we presented a risk-based KYC framework to address the increasing complexity of AML compliance for FIs and SMBs. As part of this framework, we used SimRank to address the challenge of identifying known associates of PEPs, closing critical gaps in contemporary compliance practices. We also published a GPU accelerated implementation of the SimRank algorithm. We leveraged our findings from studying UBO registry in Luxembourg to introduce two risk indicators, (1) involvement with underage shareholders, and (2) number of companies at the address. These two indicators, as well as others inspired by AML directives, were combined into a single modular risk metric that can be used to decide whether it is necessary to perform EDD. We then performed experiments to evaluate our approach in 4 different scenarios. We also discussed how future work can address some of the issues that were not covered in this work. Lastly, we are contributing to the research community by releasing an anonymized version of the UBO dataset. This release will enable fellow researchers to validate, replicate, or extend our work, promoting transparency and collaboration in the ongoing development of AML compliance tools. Bridging technology and compliance is essential to address the challenges posed by the increasingly interconnected and complex global financial system.

## Acknowledgment

This research was funded in whole, or in part, by the Luxembourg National Research Fund (FNR), grant reference NCER22/IS/16570468/NCER-FT. For the purpose of open access, and in fulfilment of the obligations arising from the grant agreement, the author has applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission. Some of the experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg [18] – see https://hpc.uni.lu

## References

[1] United Nations Office on Drugs and Crime, "Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes," https://www.unodc.org/documents/data-and-analysis/Studies/Illicit-financial-flows_31Aug11.pdf, 2011, accessed: 2023-11-03.

[2] Europol, "Criminal asset recovery in the eu," https://www.europol.europa.eu/cms/sites/default/files/documents/criminal_asset_recovery_in_the_eu_web_version.pdf, 2016, accessed: 2023-11-06.

[3] J. C. Sharman, "Shopping for Anonymous Shell Companies: An Audit Study of Anonymity and Crime in the International Financial System," *Journal of Economic Perspectives*, vol. 24, no. 4, pp. 127–140, Dec. 2010.

[4] E. van der Does de Willebois, E. M. Halter, R. A. Harrison, J. W. Park, and J. C. Sharman, *The Puppet Masters : How the Corrupt Use Legal Structures to Hide Stolen Assets and What to Do About It*. World Bank, 2011.

[5] A. N. Eddin, J. Bono, D. Aparício, D. Polido, J. T. Ascensão, P. Bizarro, and P. Ribeiro, "Anti-money laundering alert optimization using machine learning with graphs," *CoRR*, vol. abs/2112.07508, 2021.

[6] Z. Tang, H. E, M. Sun, L. Zhao, R. Wang, and M. Song, "Anti-money laundering method based on hierarchical risk control knowledge graph," in *International Conference on Artificial Intelligence and Computer Science*, 2023.

[7] B. Dumitrescu, A. Băltoiu, and S. Budulan, "Anomaly detection in graphs of bank transactions for anti money laundering applications," *IEEE Access*, vol. 10, pp. 47 699–47 714, 2022.

[8] A. K. Shaikh, M. Al-Shamli, and A. Nazir, "Designing a relational model to identify relationships between suspicious customers in anti-money laundering (aml) using social network analysis (sna)," *Journal of Big Data*, vol. 8, 2021.

[9] G. Jeh and J. Widom, "Simrank: a measure of structural-context similarity," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 538–543.

[10] Court of Justice of the European Union. (2023) Judgment of 22 november 2022, joined cases c-37/20 and c-601/20. [Online]. Available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=268059&doclang=en

[11] Organized Crime and Corruption Reporting Project (OCCRP), "Boss babies: The children who own hundreds of luxembourg corporations," OCCRP, 2021, accessed: 2023-11-03. [Online]. Available: https://www.occrp.org/en/openlux/boss-babies-the-children-who-own-hundreds-of-luxembourg-corporations

[12] International Consortium of Investigative Journalists (ICIJ), "The paradise papers," Online, 2017, available: https://www.icij.org/investigations/paradise-papers/.

[13] ——, "The panama papers," Online, 2016, available: https://www.icij.org/investigations/panama-papers/.

[14] Financial Action Task Force (FATF), "High-risk and other monitored jurisdictions," Financial Action Task Force, 2023, accessed: 2023-11-03. [Online]. Available: https://www.fatf-gafi.org/en/countries/black-and-grey-lists.html

[15] Transparency International, "Corruption perception index," Transparency International, 2023, accessed: 2023-11-03. [Online]. Available: https://www.transparency.org/en/cpi/2022

[16] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *CoRR*, vol. abs/2005.11401, 2020.

[17] M. A. Naheem, "Money laundering using investment companies," *Journal of Money Laundering Control*, vol. 18, no. 4, pp. 438–446, Jan. 2015, publisher: Emerald Group Publishing Limited.

[18] S. Varrette, H. Cartiaux, S. Peter, E. Kieffer, T. Valette, and A. Olloh, "Management of an Academic HPC & Research Computing Facility: The ULHPC Experience 2.0," in *Proc. of the 6th ACM High Performance Computing and Cluster Technologies Conf. (HPCCT 2022)*. Fuzhou, China: Association for Computing Machinery (ACM), July 2022.