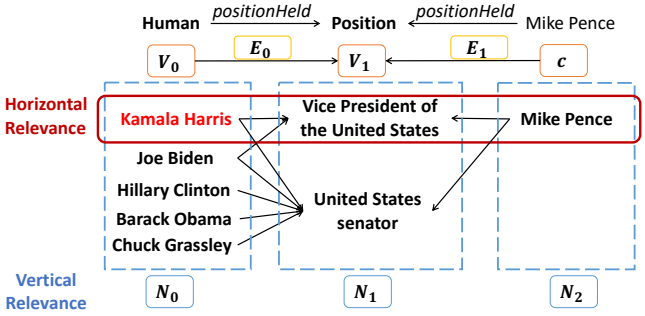

Figure 1: An example for definitions on relevance.

sets N_0, N_1 and N_2 as depicted in Figure 1. Taking N_1 as an example, its proximity to Pence is calculated as:

$$S(Pence, N_1) = \frac{1}{2}[S(Pence, VicePresident) + S(Pence, senator)].$$

The HR score is taken as the maximum PR (in red box), which is further calculated as the minimum among: $S(Pence, Harris)$, $\beta^{-1} \cdot S(Pence, VicePresident)$ and $\beta^{-2} \cdot S(Pence, Pence)$. The VR score is calculated as the minimum among: $S(Pence, N_0)$, $\beta^{-1} \cdot S(Pence, N_1)$ and $\beta^{-2} \cdot S(Pence, N_2)$.

2.3 Strikingness Measure

Given a context-anchored pattern, we can derive the context-aware peer entities. Furthermore, we extract OFs that distinguish the target entity from its peer entities. In this section, we present the definitions of *candidate OF* and *strikingness measure* for the OF extraction.

Definition 9 (Candidate OF). *A candidate OF is defined as a quadruple $Q = (t, \mathcal{A}, \mathcal{X}, P(\tilde{v_0}, \tilde{v_\ell} = c))$ with $|N_0| \geq w$, where the symbol meanings are listed as follows.*

- *t, c: The target and context nodes in $\mathcal{G}(\mathcal{V}, \mathcal{E})$.*
- *$\mathcal{A}$: An attribute of t.*
- *$\mathcal{X}$: The value for attribute $\mathcal{A}$ of t.*
- *$P(\tilde{v_0}, \tilde{v_\ell} = c)$: An ℓ-hop context-anchored pattern, which describes the relationship under consideration.*
- *w: A significance threshold to ensure that the number of peer entities is large enough for an interesting OF.*

A candidate OF comprises an attribute-value pair of the target entity t, which is striking if it makes the target stand out compared to its peer entities. To identify an OF, we need to rank the attribute-value pairs of the target against its peer entities according to a strikingness measure. We adopt a strikingness measure from [30, 35], which identifies outstanding attribute-value pairs utilizing statistics on the attribute-value frequency distribution. Given a candidate OF $Q = (t, \mathcal{A}, \mathcal{X}, P(\tilde{v_0}, \tilde{v_\ell} = c))$, $F(\mathcal{A}, \mathcal{X}', N_0)$ denotes the **frequency** (percentage) of a value $\mathcal{X}'$ among its peer entities in N_0, i.e., the matching node set of $\tilde{v_0}$.

Definition 10 (Strikingness Measure). *The strikingness score of a candidate OF $Q(t, \mathcal{A}, \mathcal{X}, P(\tilde{v_0}, \tilde{v_\ell} = c))$ is measured as below:*

$$\mathcal{I}(Q) = \sum_{\mathcal{X}' \in \overline{\mathcal{X}}} F(\mathcal{A}, \mathcal{X}', N_0),$$

where $\overline{\mathcal{X}} = \{\mathcal{X}' | F(\mathcal{A}, \mathcal{X}', N_0) > F(\mathcal{A}, \mathcal{X}, N_0)\}$.

Table 1: An example to illustrate the calculation of the strikingness score for Figure 1.

Attribute	Value	Entity	Frequency
Gender	Female	Kamala Harris	0.4
Gender	Female	Hillary Clinton	
Gender	Male	Joe Biden	0.6
Gender	Male	Barack Obama	
Gender	Male	Chuck Grassley	

By Definition 10, a fact is more striking if its value for the attribute $\mathcal{A}$ of the target entity t is *rarer*, i.e. has lower frequency, than *most* others. While alternative strikingness measures [3, 18, 25, 35] exist and could be used with our design. We have adopted an existing measure that has been shown to be effective in finding OFs. Example 3 provides a toy example for the strikingness scoring.

Example 3. *From Figure 1, a candidate OF is Q = (Harris, gender, Female, P(Human, position, Pence)). In Table 1, the value **Male** for attribute **gender** has a strictly higher frequency than **Female**. Thus, $\mathcal{I}(Q)$ sums to 0.6, as there are no other values to be considered.*

2.4 The COF Mining Problem

We formalize the mining process as the top-(k, l) COF problem in Definition 11. The context-awareness of a COF is achieved by ensuring the relevance of *context-aware peer entities* as well as their relationships, i.e., *context-anchored patterns*, associated with the context entity.

Definition 11 (Top-(k,l) COF Problem). *Given the target and context $\langle t, c \rangle$, and $\mathcal{G}(\mathcal{V}, \mathcal{E})$, we find the top-$l$ COFs with the highest strikingness scores from the top-k relevant context-anchored patterns.*

To solve the top-(k, l) COF problem, the processing divides into two steps. In the first step, we find the top-k relevant context-anchored patterns ranked by $\mathcal{R}(\cdot)$ in Definition 8. In the second step, we extract the top-l COFs from the collected k patterns. The major challenge lies in the first step, as the second step can be based on existing well-developed approaches [30, 35]. In the first step, in order to identify context-anchored patterns, we need to first conduct path enumeration to find connecting paths between the target and the context nodes. Then, for each connecting path, we need to enumerate all possible patterns, and for each enumerated pattern, we further find all its matching instances in the graph for calculating the relevance score. Every procedure involved here is of exponential complexity. Even worse, the scale of today's open KGs renders the problem more difficult. To mitigate the efficiency issue, we propose optimizations to quickly prune unpromising patterns and avoid any redundant computation to speed up the search. As a result, we reduce the search time from more than tens of seconds to sub-seconds on average over hundreds of queries, running on a KG with more than a half billion edges. Before moving on, we refer readers to Table 2 for the frequently used notations.

3 ALGORITHMS AND OPTIMIZATIONS

In this section, we first introduce the baseline algorithm. Then, we present several optimizations that speed up the processing.

3.1 Overview and Baseline Algorithm

Given $\langle t, c \rangle$, the baseline algorithm runs in two steps to produce the top-(k, l) COFs. In the **first step**, we traverse the graph with a

Table 7: Detailed output. $(\mathcal{A}, \mathcal{X}, \mathcal{I}, |N_0|)$ denotes the attribute, value, strikingness score, and the number of context-aware peer entities. NL is short for Natural Language. We use E^{-1} to denote a right-to-left edge direction.

	Maverick	FMiner
Case 1	⟨ **Michelle Obama**, **Hillary Clinton**⟩	
Path Pattern	P(**Human**, $familyName$, Robinson)	P(**Human**, $positionHeld$, **FirstLadyOfUS**, $positionHeld^{-1}$, Hillary Clinton)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($familyName$, Obama, 0.999519, 2086)	($positionHeld$, Dean, 0.948276, 58)
NL	Among all people who have family name Robinson, Michelle is the only that also has Obama as her family name.	Among the first ladies of the US, Michelle is the only one who has been Dean in any university.
Case 2	⟨**Akon**, **Michael Jackson** ⟩	
Path Pattern	P(**Human**, $occupation$, actor)	P(**Human**, $genre$, **MusicGenre**, $influencedBy^{-1}$, Michael Jackson)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($occupation$, philanthropist, 0.999616, 190470)	($countryOfCitizenship$, Senegal, 0.996012, 7812)
NL	Among all actors, Akon is one of a few philanthropists.	Among people whose music genre is influenced by Michael Jackson, Akon is the only Senegalese-American.
Case 3	⟨**Steve Jobs**, **Bill Gates** ⟩	
Path Pattern	P(**Human**, $occupation$, engineer)	P(**Human**, $occupation$, **inventor**, $occupation^{-1}$, Bill Gates)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($occupation$, executive producer, 0.999475, 28622)	($occupation$, executive producer, 0.995956, 5193)
NL	Among all engineers, Steve Jobs is one of a few executive producers.	Among all inventors, Steve Jobs is the only executive producer.
Case 4	⟨**Lionel Messi**, **Neymar** ⟩	
Path Pattern	P(**Human**, $participateIn$, 2008 Summer Olympics)	P(**Human**, $memberOf$, **F.C. Barcelona**, $memberOf^{-1}$, Neymar)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($participateIn$, 2019 Copa América, 0.999203, 10391)	($memberOf$, Argentina national football team, 0.986425, 1547)
NL	Among all participants of 2008 Summer Olympics, Messi is one of a few who also participate in 2019 Copa América.	Among all team members of F.C. Barcelona, Messi is one of a few that also play in Argentina national football team.
Case 5	⟨**Donald Trump**, **Joe Biden** ⟩	
Path Pattern	P(**Human**, $occupation$, politician)	P(**Human**, $awardReceived$, **Medal**, $awardReceived^{-1}$, Joe Biden)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($occupation$, game show host, 0.999978, 462240)	($awardReceived$, WWE Hall of Fame, 0.966346, 623)
NL	Among all politicians, Trump is one of a few game show hosts.	Among all people receiving some same medal as Joe Biden, Trump is the only one who has received the award of WWE Hall of Fame.
Case 6	⟨**Barack Obama**, **Donald Trump**⟩	
Path Pattern	P(**Human**,$occupation$, politician)	P(**Human**, $positionHeld$, **USPresident**, $positionHeld^{-1}$, Donald Trump)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($occupation$, Community Organizer, 0.999948, 462240)	($language$, Indonesian, 0.977273, 46)
NL	Among all politicians, Obama is one of a few community organizers.	Among all the presidents of US, Obama is the only one who can speak Indonesian.
Case 7	⟨**Taylor Swift**, **Lady Gaga**⟩	
Path Pattern	P(**Human**, $occupation$, actor)	P(**Human**, $Genre$, **MusicGenre**, $Genre^{-1}$, Lady Gaga)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($occupation$, banjoist, 0.999685, 190470)	($instrument$, banjo, 0.994737, 220)
NL	Among all actors, Taylor is one of a few banjoists.	Among people with the same music genre as Lady Gaga, Taylor is the only one whose instrument includes banjo.
Case 8	⟨**Apple Inc.**, **Microsoft** ⟩	
Path Pattern	P(**Business**, $industry$, **Industry**, $industry^{-1}$, GOG.com)	P(**Business**, $industry$, **software industry**, $industry^{-1}$, Microsoft)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($industry$, consumer electronics, 0.999206, 3778)	($industry$, consumer electronics, 0.956522, 92)
NL	Among all businesses that share a same industry as GOG.com, Apple Inc. is one of a few that belong to the consumer electronics industry.	Among all businesses that fall in software industry, Apple Inc. is one of few that also belong to the consumer electronics industry.
Case 9	⟨**Jason Statham**, **Vin Diesel** ⟩	
Path Pattern	P(**Human**, $castMemberOf$, **Film**, $distributor$, Universal Pictures)	P(**Human**, $genre$, action movie, $genre^{-1}$, Vin Diesel)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($occupation$, competitive diver, 0.987717, 13523)	($industry$, kickboxer, 0.960000, 50)
NL	Among all people starring in any films of Universal Pictures, Jason Statham is the only competitive diver.	Among all people whose movie genre is action movie, Jason Statham is the only kickboxer.
Case 10	⟨**LeBron James**, **Michael Jordan** ⟩	
Path Pattern	P(**Human**, $positionPlayedOnTeam$, **BasketballPosition**, $sport$, basketball)	P(**Human**, $awardReceived$, **NBA RookieOfTheYearAward**, $awardReceived^{-1}$, Michael Jordan)
$(\mathcal{A}, \mathcal{X}, \mathcal{I}, \lvert N_0\rvert)$	($positionPlayedOnTeam$, point forward, 0.999752, 20191)	($positionPlayedOnTeam$, point forward, 0.985915, 71)
NL	Among all people playing in some basketball team, Lebron James is one of few who play point forward.	Among all people who won the NBA Rookie of the year award as Michael Jordan, Lebron James is one of few who play point forward.

In Table 7, we show all details of output facts used in the user study. Note that only the fact descriptions are shown to participates so that participants are not overwhelmed. The target and context entities are only mentioned in the questions associated with each user study case. From the details shown in Table 7, Maverick is capable of finding OFs with higher strikingness scores than FMiner. The strikingness measure favors OFs that are generated from a large number of peer entities, which can result in higher strikingness scores. This can make the OFs sound more striking on one hand, e.g. Case 2 and 7 where Maverick outperforms FMiner in Figure 5. On the other hand, it sometimes can lead to less meaningful OFs, e.g. Case 1 and 6 where FMiner performs better.

From the results, the peer entities generated by FMiner do not have a large cardinality as Maverick. FMiner effectively confines the OF extraction to the relevant context, though losing some strikingness scores. There are some very interesting results, such as Case 1 where FMiner finds all first ladies of US, Case 6 where FMiner identifies all the US presidents, and Case 9 where FMiner navigates to the action movie genre that is very relevant to the two involved actors.

We note that FMiner relies on the accuracy and completeness of the underlying KG. Fortunately, Wikidata is rapidly growing and its content is covering more and more entities as well as topics. This makes COF finding from such a KG even more attractive.