CoReRank: Ranking to Detect Users Involved in Blackmarket-based Collusive Retweeting Activities (Supplementary Material)

1 Parameter Selection

We conducted Parameter Selection by the method of Parameter Sweep. In Parameter Sweep, all possible combinations of the parameters are checked to find out the most optimal combination that yields the best precision.

Selecting Graph Parameters

The weights of the edges of the graph, w_r and w_q are taken in the range $(0, 1)$ at a step of 0.25. Also, the condition $0 < w_r \leq w_q < 1$ needs to be satisfied. Thus, all the possible combinations of (w_r, w_q) are ordered in the following fashion- $(0.25, 0.25)$, $(0.25, 0.5)$, $(0.25, 0.75)$, $(0.5, 0.5)$, $(0.5, 0.75)$ and $(0.75, 0.75)$.

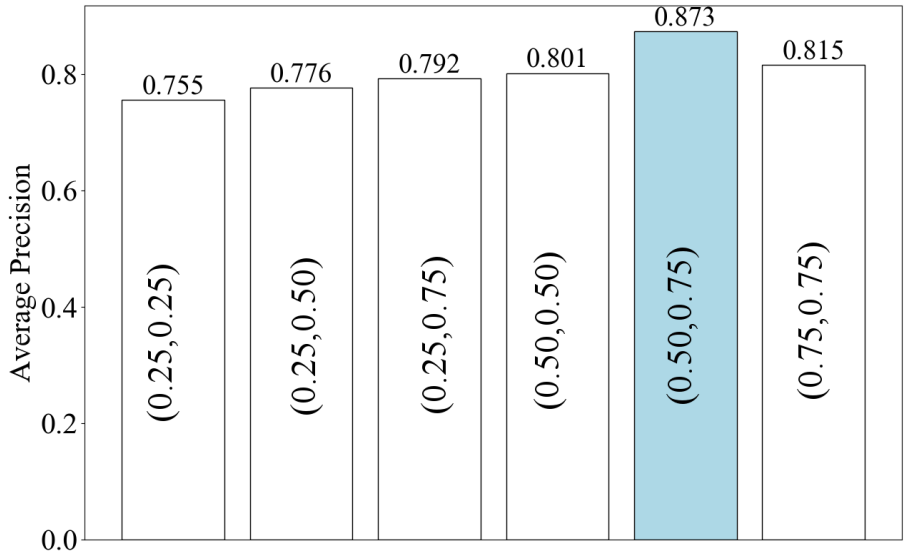

Figure 1: **Average precision of CoReRank vs. Index of ordered combinations of edge weights**

Figure 1 shows how the precision of the algorithm varies with the different ordered combinations of the weights of the edges of the graph. The highest precision is observed at $w_r = 0.50$ and $w_q = 0.75$, which is the optimal parameter combination fixed for our entire experiment.

Selecting Recurrence Parameters

CoReRank has 7 parameters - $\gamma_{1t}, \gamma_{2t}, \gamma_{3t}, \gamma_{1u}, \gamma_{2u}, \gamma_{3u}, \gamma_{4u}$.
$\gamma_{1t}, \gamma_{2t}, \gamma_{3t}, \gamma_{1u}, \gamma_{2u}, \gamma_{4u}$ are taken in the range $[0, 1]$ with a step of 0.3, while γ_{3u} is chosen from $\{0, 1, 2, 3\}$.

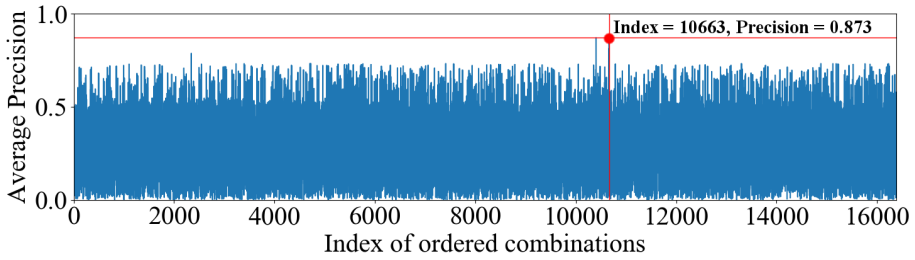

Figure 2: **Average precision of CoReRank vs. Index of ordered combinations of parameters**

Figure 2 demonstrates how the precision changes due to the affect of different combinations of the parameters. The `x-axis` denotes the combinations of the parameters. The combinations are ordered, in order to facilitate the plotting of the curve. The `y-axis` denotes the precision of the algorithm. As it can be noticed, the precision of the algorithm reaches a global maxima at index 10663. The combination that is represented by this index 10663 is selected as our optimal parameter combination throughout all the experiments. This combination is as follows - $\gamma_{1t} = 0.6$, $\gamma_{2t} = 0.6$, $\gamma_{3t} = 0.3$, $\gamma_{1u} = 0.6$, $\gamma_{2u} = 0.6$, $\gamma_{3u} = 3$ and $\gamma_{4u} = 0.3$.

2 Proofs for lemma and theorems

This supplementary sheet contains the complete proofs of the lemma and theorems discussed in the submitted paper.

Before we begin, here are the formulas for the credibility scores and merit scores used in the algorithm to calculate scores for the k^{th} iteration:

$$\tilde{C}^{k-1}(u) = \texttt{norm}(C^{k-1}(u)) \forall u \in U \; such \; that \; \tilde{C}^{k-1}(u) \in [0, 1] \tag{1}$$

$$M^k(t) = \frac{\sum_{u \in \text{In(t)}} \gamma_{1t} \cdot \tilde{C}^{k-1}(u) \cdot S(u,t) + \gamma_{2t} \cdot \pi_T(t) + \gamma_{3t} \cdot \mu_T}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|} \tag{2}$$

$$C^k(u) = \frac{\sum_{t \in \text{Out(u)}} \gamma_{1u} \cdot M^k(t) \cdot S(u,t) + \gamma_{2u} \cdot \pi_U(u) + +\gamma_{3u} \cdot \tau_u + \gamma_{4u} \cdot \mu_U}{\gamma_{1u} + \gamma_{2u} + \gamma_{3u} + \gamma_{4u} + |\text{Out(u)}|} \tag{3}$$

where $\gamma_{1t}, \gamma_{2t}, \gamma_{3t}, \gamma_{1u}, \gamma_{2u}, \gamma_{3u}, \gamma_{4u}$ are constants provided as inputs to the algorithm. Also, it is important to note that $S(u,t) \leq \frac{3}{4}$, as maximum weight of an edge can be $w_q = \frac{3}{4}$

[**Lemma 1**] For any given tweet, t, the difference between their final score and their score after the first iteration of cannot exceed $\frac{3}{4}$. Formally, it means that $|M^\infty(t) - M^1(t)| \leq \frac{3}{4}$. Similarly, for users the upper bound is $\frac{3}{4}$, i.e., $|C^\infty(u) - C^1(u)| \leq \frac{3}{4}$

Proof.

Let us first prove that $|M^\infty(t) - M^1(t)| \leq \frac{3}{4}$.
From 2, we know that when the algorithm converges,

$$M^\infty(t) = \frac{\sum_{u \in \text{In(t)}} \gamma_{1t} \cdot \tilde{C}^\infty(u) \cdot S(u,t) + \gamma_{2t} \cdot \pi_T(t) + \gamma_{3t} \cdot \mu_T}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|}$$

Similarly after the first iteration,

$$M^1(t) = \frac{\sum_{u \in \text{In(t)}} \gamma_{1t} \cdot \tilde{C}^0(u) \cdot S(u,t) + \gamma_{2t} \cdot \pi_T(t) + \gamma_{3t} \cdot \mu_T}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|}$$

Hence, substituting values in $|M^\infty(t) - M^1(t)|$, we get,

$$|M^\infty(t) - M^1(t)| = |\frac{\sum_{u \in \text{In(t)}} \gamma_{1t} \cdot (\tilde{C}^\infty(u) - \tilde{C}^0(u)) \cdot S(u,t)}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|}|$$

Since $|x + y| \leq |x| + |y|$,

$$|M^\infty(t) - M^1(t)| \leq \frac{\sum_{u \in \text{In(t)}} |\gamma_{1t} \cdot (\tilde{C}^\infty(u) - \tilde{C}^0(u)) \cdot S(u,t)|}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|}$$

As $|x \cdot y| = |x| \cdot |y|$,

$$|M^\infty(t) - M^1(t)| \leq \frac{\sum_{u \in \text{In(t)}} |\gamma_{1t}| \cdot |(\tilde{C}^\infty(u) - \tilde{C}^0(u))| \cdot |S(u,t)|}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|}$$

Since $\tilde{C}^k(u) \in [0,1] \forall k \in [0,\infty), \forall u \in U$, $|(\tilde{C}^\infty(u) - \tilde{C}^0(u))| \leq 1$.
Also, $S(u,t) \leq \frac{3}{4}$ and $\gamma_{1t} \in [0,1)$

$$\therefore |M^\infty(t) - M^1(t)| \leq \frac{\sum_{u \in \text{In(t)}} \gamma_{1t} \cdot 1 \cdot \frac{3}{4}}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|}$$

$$\therefore |M^{\infty}(t) - M^{1}(t)| \leq \frac{\gamma_{1t} \cdot |In(t)| \cdot \frac{3}{4}}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|} \leq \frac{3}{4}$$

Similarly, it is possible to show that $|C^{\infty}(u) - C^{1}(u)| \leq \frac{3}{4}$

[**Theorem of convergence**] Between successive iterations, the difference in the scores of the users and tweets is bounded. For any user $u \in U$, $|C^{\infty}(u) - C^{k}(u)| \leq \left(\frac{3}{4}\right)^{k}$. Thus, as the algorithm proceeds through more and more iterations, the value of k keeps on increasing and the difference in score from the final score keeps on decreasing. Similarly for a tweet $t \in T$, $|M^{\infty}(t) - M^{k}(t)| \leq \left(\frac{3}{4}\right)^{k-1}$.

Proof.
We will prove these using induction on k.
Base Cases ($k = 1$): We know from Lemma 2, $\forall u \in U, \forall t \in T$

$$|C^{\infty}(u) - C^{1}(u)| \leq \frac{3}{4}$$

and

$$|M^{\infty}(t) - M^{1}(t)| \leq \frac{3}{4}$$

Hence, the base cases are satisfied.
Induction Hypothesis: Assume that for any $n \leq k$, we have ,

$$|C^{\infty}(u) - C^{k}(u)| \leq (\frac{3}{4})^{k}$$

and

$$|M^{\infty}(t) - M^{k}(t)| \leq (\frac{3}{4})^{k-1}$$

$\forall u \in U, \forall t \in T$.

Induction Step ($k = n + 1$): For $k = n + 1$, we have,

$$|M^{\infty}(t) - M^{n+1}(t)| = |\frac{\sum_{u \in \text{In(t)}} \gamma_{1t} \cdot (\tilde{C}^{\infty}(u) - \tilde{C}^{n}(u)) \cdot S(u,t)}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|}|$$

$$\implies |M^{\infty}(t) - M^{n+1}(t)| \leq \gamma_{1t} \frac{\sum_{u \in \text{In(t)}} |(\tilde{C}^{\infty}(u) - \tilde{C}^{n}(u))| \cdot |S(u,t)|}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|}$$

As $|(\tilde{C}^{\infty}(u) - \tilde{C}^{n}(u))| \leq |C^{\infty}(u) - C^{n}(u)| \leq \left(\frac{3}{4}\right)^{n}$,

$$\implies |M^{\infty}(t) - M^{n+1}(t)| \leq \frac{\gamma_{1t} \cdot \left(\frac{3}{4}\right)^{n} \cdot |\text{In(t)}| \cdot |S(u,t)|}{\gamma_{1t} + \gamma_{2t} + \gamma_{3t} + |\text{In(t)}|} \leq (\frac{3}{4})^{n}$$

Thus, $\forall t \in T$, $|M^{\infty}(t) - M^{k}(t)| \leq \left(\frac{3}{4}\right)^{k-1}$ by proof of induction.
Similarly,

$$|C^{\infty}(u) - C^{n+1}(u)| = |\frac{\sum_{t \in \text{Out(u)}} \gamma_{1u} \cdot (M^{\infty}(t) - M^{n+1}(t)) \cdot S(u,t)}{\gamma_{1u} + \gamma_{2u} + \gamma_{3u} + \gamma_{4u} + |\text{Out(u)}|}|$$

$$\implies |C^{\infty}(u) - C^{n+1}(u)| \leq \frac{\sum_{t \in \text{Out(u)}} |\gamma_{1u}| \cdot |(M^{\infty}(t) - M^{n+1}(t))| \cdot |S(u,t)|}{\gamma_{1u} + \gamma_{2u} + \gamma_{3u} + \gamma_{4u} + |\text{Out(u)}|}$$

$$\implies |C^{\infty}(u) - C^{n+1}(u)| \leq \frac{|\gamma_{1u}| \cdot \left(\frac{3}{4}\right)^n \cdot |S(u,t)|}{\gamma_{1u} + \gamma_{2u} + \gamma_{3u} + \gamma_{4u} + |\text{Out(u)}|}$$

Since $S(u,t) \leq \frac{3}{4}$,

$$\implies |C^{\infty}(u) - C^{n+1}(u)| \leq \frac{|\gamma_{1u}| \cdot \left(\frac{3}{4}\right)^{n+1}}{\gamma_{1u} + \gamma_{2u} + \gamma_{3u} + \gamma_{4u} + |\text{Out(u)}|} \leq \left(\frac{3}{4}\right)^{n+1}$$

Thus, $\forall u \in U$, $|C^{\infty}(u) - C^{k}(u)| \leq \left(\frac{3}{4}\right)^{k}$ by proof of induction.
This completes our proof.

[**Bound on iterations**] There exists a bound on the number of iterations until converges. This bound is governed by the precision to which the score is calculated before convergence is declared, i.e., ϵ. The number of iterations till convergence is at most $2 + \lceil \frac{\log\left(\frac{\epsilon}{2}\right)}{\log\left(\frac{3}{4}\right)} \rceil$

Proof: Let $k = 2 + \lceil \frac{\log\left(\frac{\epsilon}{2}\right)}{\log\left(\frac{3}{4}\right)} \rceil$. By Theorem 2, after $k+1$ iterations, $\forall t \in \mathcal{T}$, $|M^{\infty}(t) - M^{k+1}(t)| \leq \frac{3}{4}^{k} \leq \frac{3}{4}^{\log_{\frac{3}{4}}(\frac{\epsilon}{2})} = \frac{\epsilon}{2}$. Similarly, $M^{\infty}(t) - M^{k+2}(t)| \leq \frac{\epsilon}{2} \cdot \frac{3}{4} \leq \frac{\epsilon}{2}$. Thus,

$$|M^{k+1}(t) - M^{k+2}(t)| = |M^{k+1}(t) - M^{\infty}(t) + M^{\infty}(t) - M^{k+2}(t)|$$

As $|x+y| \leq |x| + |y|$,
$\implies |M^{k+1}(t) - M^{k+2}(t)| \leq |M^{k+1}(t) - M^{\infty}(t)| + |M^{\infty}(t) - M^{k+2}(t)| \leq 2 \cdot \frac{\epsilon}{2} = \epsilon$
Similarly for credibility, we have, $|C^{k+1}(u) - C^{k+2}(u)| \leq \epsilon \forall u \in \mathcal{U}$.
Thus, by line 6 of Algorithm 1 it will take $k+2$ iterations to converge.

3 Thresholding for Semi-supervised evaluations

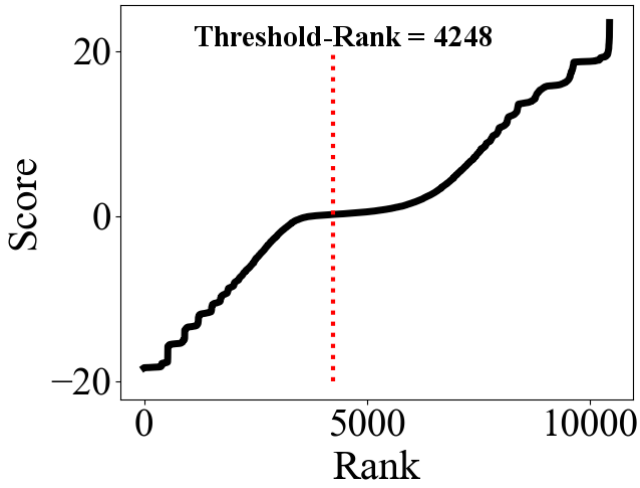

Figure 3: **Non-cumulative distribution of credibility scores vs. rank of users after algorithm**

In order to evaluate CoReRank+, we create a thresholding technique that allows us to compare it's results with the available ground truth labels.

In Figure 3, we plotted the non-cumulative distribution of credibility scores for all users ($C(u) \forall u \in U$) against their allotted ranks. At $x = 4248$, we encounter the **Sharpest Turning Point** of the curve. Corresponding to $x = 4248$, $C(u) = 0.202$. A threshold margin is then created at $x = 4248$. All the users whose rank lies before 4248 are termed as `collusive` and all those whose rank is after 4248 are termed as `genuine`.