Using Causality to Infer Coordinated Attacks in Social Media

Anonymous submission

Abstract

The rise of social media has been accompanied by a dark side with the ease of creating fake accounts and disseminating misinformation through coordinated attacks. Existing methods to identify such attacks often rely on thematic similarities or network-based approaches, overlooking the intricate causal relationships that underlie coordinated actions. This work presents an approach for detecting coordinated attacks using Convergent Cross Mapping (CCM), a technique that infers causality from temporal relationships between user activity. We apply CCM to real-world data from the infamous IRA attack on US elections, achieving F1 scores up to 75.3% in identifying coordinated accounts. Our results demonstrate the effectiveness of our model in uncovering causal structures of coordinated behavior, offering a promising avenue for mitigating the threat of malicious campaigns on social media platforms.

1 Introduction

While social media platforms have witnessed explosive growth due to factors like peer pressure, evolving communities, and influencer culture, this increased engagement has fueled a parallel threat: the ease of generating fake accounts has increased the spread of misinformation and disinformation. Politically driven campaigns, seeking to manipulate public opinion and achieve specific goals, rely on large numbers of coordinated accounts to amplify their messages and maximize the impact. Existing techniques primarily focus on identifying coordinated behaviors based on thematic similarities, overlooking the intricate causality relationships that underlie coordinated actions. This complexity in detecting true coordination represents a significant gap in current methodologies, forming the core motivation for our work.

There have been numerous examples of coordinated attacks on social media. One of the most notable instances was the influence of Russia's IRA on the USA presidential elections via Twitter and Facebook (Mueller, Internet Research Agency, and States 2018). The Permanent Select Committee on Intelligence identified 3,841 coordinated Twitter accounts and 470 Facebook pages that were affiliated with the IRA in 2017. In 2018, Twitter publicly released tweets and users related to this case. In 2019, the UK general elections were influenced by coordinated users who polarized political opinions on Twitter (Nizzoli et al. 2021).

There have been numerous attempts to identify coordination in Online Social Networks (OSNs). Some work in this field sought to identify campaigns in social media (Lee et al. 2011, 2012). A major limitation of their work is the assumption that coordination is reflected in the "theme" of messages while other aspects of behaviour are ignored. Network-based approaches (Pacheco et al. 2021; Nizzoli et al. 2021; Weber and Neumann 2021; Magelinski, Ng, and Carley 2021; Hristakieva et al. 2022) tend to define coordination in terms of community detection on user similarity graphs. Weber and Neumann (2021) highlighted several coordination strategies: *pollution* – flooding a community with repeated content, *boost* – heavily reposting content to make the topic appear popular, and *bully* – groups collectively harassing another individual or a community. In contrast, Zhang, Sharma, and Liu (2021) and Sharma et al. (2021) define coordination in terms of the synchronicity of users over time. They try to identify coordinated users using masked self-attention (Vaswani et al. 2017) to encode the event history, using an approach similar to the prediction model for marked temporal point processes by Shchur, Biloš, and Günnemann (2020). Network-based and theme-based approaches work under the assumption that the content is the governing factor of the coordinated behaviour. The activity-based approaches work under the assumption that active times of coordinated users are causally linked with each other. In contrast, we consider that coordination should be reflected in the user activity traces.

We propose to identify how influence flows within a community of users by assessing causality between pairs of users by exploring the layered dynamics and dependencies between users. Causality offers a nuanced understanding of the users who influence or trigger coordinated responses from others. The idea of causality not only enhances the precision of coordination detection, but also provides a deeper understanding of the mechanisms driving coordinated activities.

In order to address the problem of inferring causality between users as a basis for identifying coordination, we build on the theory of Convergent Cross Mapping (CCM) (Sugihara et al. 2012). CCM is a powerful technique that has been used to identify causality in applicaitons such as ecology and climatology. However, utility of CCM for inferring coordinated behaviour among social media users has not been considered in the literature.

In this article, we investigate methods of identifying coordination using convergent cross mapping, and evaluate the performance of our model on real data. Our research aims to address the following questions:

RQ1. How effective is inferring coordination using causal structures of users?

RQ2. Can our model be optimized further for enhanced performance in identifying coordinated attacks?

RQ3. What are the key limitations and challenges associated with inferring coordination using causal structures?

Our experiments on the IRA dataset (Permanent Select Committee on Intelligence 2018; Mueller, Internet Research Agency, and States 2018) show that cross mapping each pair of users can identify coordinated pairs of users accurately. Moreover, the coordinated users who were identified by our model belong to clearly separated clusters of interests. We achieve F1 scores up to 75.3%. Further, we exploit the clustered nature of users to optimize our model.

In Section 2, we provide the background and definitions needed for our framework and then our hypothesis and the problem statement is described. In Section 3, we describe our methodology followed by our model. The data, experiments, results, and optimizations are presented in Section 4. We further studied how to recognize leaders and uncover hidden coordinated behaviours in the same section. Finally, we conclude our study and identify directions for future research in Section 5.

2 Background and Problem Statement

Convergent Cross Mapping. Unraveling relationships within complex systems often leads researchers to study nuanced separation between correlation and causation. While correlation signifies a statistical association between two variables, it falls short of establishing a cause-and-effect relationship. In contrast, causation implies a direct influence of one variable on another, suggesting a deeper understanding of the underlying mechanisms governing a system. Convergent Cross Mapping (CCM) (Sugihara et al. 2012) is a powerful technique that can determine causality using the variation of correlation at different training sample sizes (known as library lengths) of predictions. CCM uses Takens' principle (Takens 1981) to detect if two variables belong to the same dynamic system. Consider two time series variables X and Y. CCM establishes the causality between variables by examining the predictive accuracy of a cross-mapped model built using historical Y data to reconstruct X states. Causality is suggested by the convergence of these reconstructed states towards the actual X values.

CCM has been primarily found application in ecology and (Sugihara et al. 2012; Clark et al. 2015; Ye et al. 2015; Frossard, Rimet, and Perga 2018) climatology (van Nes et al. 2015; Luo et al. 2015; Zhang, Wang, and Tsonis 2018). CCM was reviewed and provided improvements in works of Ye et al. (2015); Krakovská and Hanzely (2016) and Tsonis et al. (2018). The study in Cobey and Baskerville (2016) explores the limitations of CCM such as its sensitivity to periodicity. We consider that there is no reason for there to be such periodical fluctuations of activity of users in OSNs, but only major events govern the activity of users. CCM has not been widely studied in the context of social networks other than the work of Luo, Zheng, and Zeng (2014). However, their work is not geared to infer coordination using causal relationships, but instead to confirm their results with alternative network measures. Thereby, we identify the gap in literature that CCM has yet to be applied in social media contexts to determine coordinating behaviours.

Preliminaries. We refer to an interaction made by a user in the OSN as an *event*. For example, on Twitter (now called X), a tweet authored by a user is considered an event. The set of timestamps of events authored by a user u is called an *activity trace* $\{t_{0,u}, t_{1,u}, \dots\}$. If there is an influence flow from user u_1 to u_2, we say $u_1 \Rightarrow u_2$. If $u_1 \Rightarrow u_2$ and $u_2 \Rightarrow u_1$, we say there is a *bidirectional coupling*. If $u_1 \Rightarrow u_2$ but $u_1 \not\Rightarrow u_2$, we say that there exists a *unidirectional coupling*. We named the directed graph where the vertices are users and the edges are influence flows to be the *influence graph* for sample of users we consider.

Let U be a set of users in an OSN. Say we determine a time period $(t_{\text{start}}, t_{\text{end}})$ that presumably contains coordinated anomalous activities based on observations. Let T be the set of activity traces performed by each user in U in the above time interval.

Problem Definition. *Given a dataset of activity traces T, find pairs of users that are causally influenced unidirectionally or bidirectionally by measuring their prediction scores through Convergent Cross Mapping. Find the users that belong to such influencing pairs and mark them as coordinating users.*

3 Methodology

Motivating Example. Our results for simulated users highlight the applicability of CCM for social media data. We simulated two users u_1 and u_2 to model different stages of a simple leader-follower behavioural spectrum. We assumed that the extremes of this spectrum to be: (1) the follower can only be activated once with a lag after the leader is active, (2) the behaviour is random for both agents. We applied CCM to observe cross map prediction accuracy measured with correlation at different library lengths. Figures 1a, 1b, 1c, and 1d show our simulation results. The increasing nature of correlation when there is a leader-follower behaviour motivated us to adapt CCM to analyze on real online social network (OSN) data.

Model. Say we analyze a set of users' (U) activity in a time period $(t_{\text{start}}, t_{\text{end}})$. First, we record timestamps of events authored by each user $u \in U$ as $T_u = \{t_{0,u}, t_{1,u}, t_{2,u}, \dots\}$. Subsequently, every T_u is vectorized to a fixed size L, $X_u = \langle X_u(1), X_u(2), X_u(3), \dots, X_u(L)\rangle$. Essentially, we partition the time series into bins of size $I = {}^{(t_{\text{end}} - t_{\text{start}})}/_{L}$. Here, $X_u(a) = |\{t \mid t \in T_u, (a-1)I \leq t < aI\}|$. The embedding function $e : U \times \mathbb{Z}_{\geq 0} \rightarrow \mathbb{Z}^{E}_{\geq 0}$ transforms each time series into a series of lagged-coordinate embeddings. For a lag $\tau > 0$ and embedding size $E > 1$, a point in X_u at time t

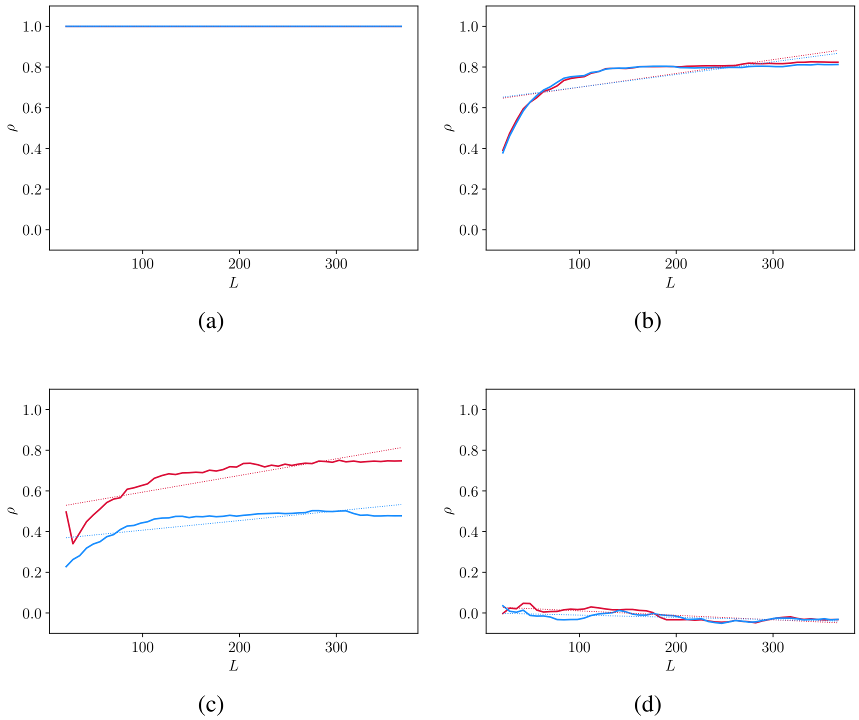

Figure 1: Motivating example of the use of CCM to model causal behaviour in simulated social media data. We show the variation of the correlation of predictions (vertical axis denoted ρ) for prediction about two simulated users u_1 and u_2, where u_2 follows u_1 on social media, as the library lengths L (i.e., sample periods) increase. CCM implies causation if the correlation is increasing for increasing library lengths. : predictions for u_1 given u_2's shadow manifold i.e., history, : predictions for u_2 given u_1's shadow manifold, : linear regression drawn for u_1's variation of correlation, and : linear regression drawn for u_2's variation of correlation. (a) u_2 posts after u_1, who posts at regular intervals. (b) u_2 posts after u_1, who posts at irregular intervals. (c) u_2 posts after u_1, who posts at irregular intervals. However, u_2 posts at random times without u_1 triggering u_2's behaviour. (d) u_1 and u_2 behaves randomly.

is transformed as $e(u,t) = \langle X_u(t), X_u(t-\tau), \ldots, X_u(t-(E-1)\tau)\rangle$. This embedding results in a manifold $M_u = [e(u,1), e(u,2), \ldots, e(u,L)]$ for each user u. For a unique pair of users u_1 and u_2, M_{u_1} and M_{u_2} can be considered as two shadow manifolds for the attractor manifold of the original behaviour system of these two users given by $M_{u_1,u_2} = [\langle X_{u_1}(t), X_{u_2}(t)\rangle \mid t = 1, 2, \ldots, L]$. We now cross map X_{u_1} using M_{u_2} and vice versa. Specifically, we use a nearest neighbors model with $k = E+1$. Unseen data in a future time window is then tested with the fitted model to obtain predictions $\hat{X}_{u_1} \mid M_{u_2}$ and $\hat{X}_{u_2} \mid M_{u_1}$. The cross-correlation ρ of each prediction $\hat{X}_{u_1} \mid M_{u_2}$ is compared with ground truth X_{u_1} for multiple library lengths. If ρ of u_1 is generally increasing with the library length, and its maximum is sufficiently great (with a threshold θ), it indicates that it is possible to estimate X_{u_1} from X_{u_2}. Therefore, in such case, we imply that the behaviour of u_1 drives u_2 (i.e., $u_1 \Rightarrow u_2$). It should also be noted that both $u_1 \Rightarrow u_2$ and $u_2 \Rightarrow u_1$ can happen at the same time. If $u_1 \Rightarrow u_2$, we mark both u_1 and u_2 to be suspected coordinated users. It is possible that u_1 is influenced by any other u_3 at different partitions of X_{u_1}. Even though it could hinder the variation of ρ, CCM successfully recovers from it since we embed only a part of history instead of the whole history. Specific hyperparameters and methodologies that are used in the submodules are given in Section 4.2.

Pairwise comparison. The computational expense associated with pairwise comparisons of N users can be substantial (${}^{N}C_2 = \mathcal{O}(N^2)$), yet accurate. In response, we devise an optimization strategy based on the observation of our raw results in Section 4.4.

4 Experiments

4.1 Data

We experiment on the dataset of the activity of Russia's Internet Research Agency (IRA) influencing the 2016 USA presidential elections (Permanent Select Committee on Intelligence 2018; Mueller, Internet Research Agency, and States 2018), which consists of confirmed coordinated activities. This is a widely used dataset for detecting coordination (Weber and Neumann 2021; Sharma et al. 2021; Zhang, Sharma, and Liu 2021; Weber and Falzon 2022) due to the availability of ground truth. The dataset consists of 8.76 mil-

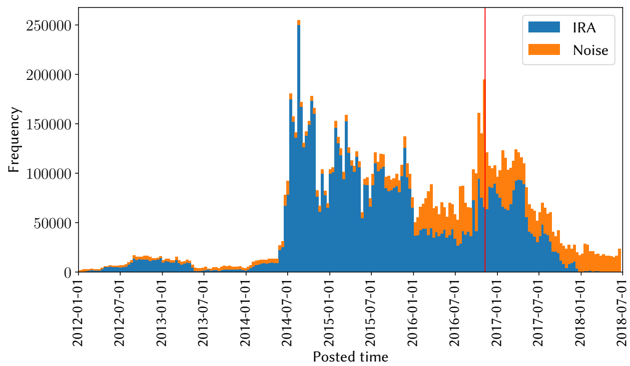

Figure 2: Stacked distribution of IRA activities and extracted noise tweets across time. The bin size for the x-axis is 1 million seconds (~11.6 days). The red vertical line shows the election date.

lion tweets posted by 3613 users. The dataset originally consisted of the following fields; Tweet id, User id, User display name, User screen name, User reported location, User profile description, User profile url, Follower count, Following count, Account creation date, Account language, Tweet language, Tweet text, Tweet time, Tweet client name, Replied tweet id, Replied user id, Quoted tweet id, Whether the tweet is a retweet, Retweeted user id, Retweeted tweet id, Latitude where the tweet is posted, Longitude where the tweet is posted, Quote count, Reply count, Like count, Retweet count, List of hashtags, List of urls, List of user mentions, List of poll choices if the tweet includes a poll. Figure 2 shows the distribution of activity across the time.

In order to test the effectiveness of a coordination detection model, we introduce a set of noisy background events to the IRA dataset, since the IRA dataset only contains the set of coordinating users. For that purpose, we scraped Twitter data for that period of time which includes the same popular hashtags in the IRA dataset using the Twitter API v2 for academics. The criteria that were used to extract data were: posted time between 2008 and 2018, marked location anywhere in the USA, contains either one of the following hashtags - *Election2016*, *MAGA*, *MakeAmericaGreatAgain*, *AmericaFirst, DonaldTrump*, *WakeUpUSA*, *Trump*, *TrumpTrain*, *HilaryClinton, Trump2016*, *DrainTheSwamp*, *TrumpPence16*, *tcot*, *POTUS*, *GOP*, *Resist*, *UniteBlue*, *NeverHillary*, *ElizabethWarren*, *WeThePeople*, *IllegalAliens*, *TrumpRussia, ImWithHer*, *GayHillary*, *WakeUpAmerica*. The above set of hashtags were the top-occurring hashtags in the original IRA dataset to ensure that the noise data belongs to the same ongoing discussions at that period of time. The background data of normal users consists of 2.80 million tweets from 333,000 of users. The distribution of coordinating tweets and the noisy tweets are shown in Figure 2. High activity is apparent near the election time period (November 2016).

4.2 Experimental Setup

Parameters. We found that a bin size I of 60 minutes and a lag τ of minimum value 1, and an embedding size E as 10 to be parameter values that yield the best results. The threshold θ was chosen as 0.5. We split the time trace vectors into 3:1 ratio for train, test datasets.

Time Intervals. The time period $(t_{\text{start}}, t_{\text{end}})$ was chosen such that it includes the election time period (November 2016) with the assumption that the coordinated activity was at a maximum during that period of time. Thus, t_{start} was chosen as July 2016 and t_{end} was chosen as November 2016.

Extracting Top Users. For the tests to be fair, we mix the top N_C number of known coordinating users and top N_N number of known normal users from the above IRA dataset. The users for tests were selected based on the frequency of activities in the testing time period in order to ensure we have enough data to cross map each and every user.

Submodules To measure correlation ρ, we use Pearson's correlation method. In order to measure the general increase in correlation values for multiple library lengths, a straightforward linear regression was conducted, and the resulting gradient was used to assert the growth.

4.3 Results

Following is a report of our results for $N_C = 200$ and $N_N = 200$. Out of ${}^{400}C_2$ number of user pairs checked, 2404 pairs were identified as coordinating pairs. Out of such pairs,

- 2319 (96.5%) were known coordinating – coordinating pairs.
- 63 (2.6%) were known coordinating – normal pairs.
- 22 (0.9%) were known normal – normal pairs.

Since we mark each user who belongs to at least one $u_1 \Rightarrow u_2$ pair as coordinating, our model detected 165 users as coordinating. For that case, the precision is 80.0% and recall is 72.0% for detecting a coordinated user out of a mix of users. The model took 642 minutes to train and predict on an M1 MacBook. Refer to Table 2 for other dataset sizes.

Figure 3a displays a graph we constructed using vertices as users and edges as influence flows identified by our model. It is apparent that there are four visible clusters of tightly coupled users for this sample. Figure 3b and 3c are described in the following subsections. There, we demonstrate how we exploited the clustered nature of users to optimize our model.

4.4 Optimizations

Motivation. In order to formally identify sub-communities in the graph in Figure 3a, we performed community detection (Blondel et al. 2008) on our results. The community detection algorithm could detect 5 main sub-communities. The colored sub-communities are shown in Figure 3b.

In a perfect scenario, say we could detect n equal sized clusters in a set of users U of size N. If we only compare user pairs within the clusters, our search space is reduced

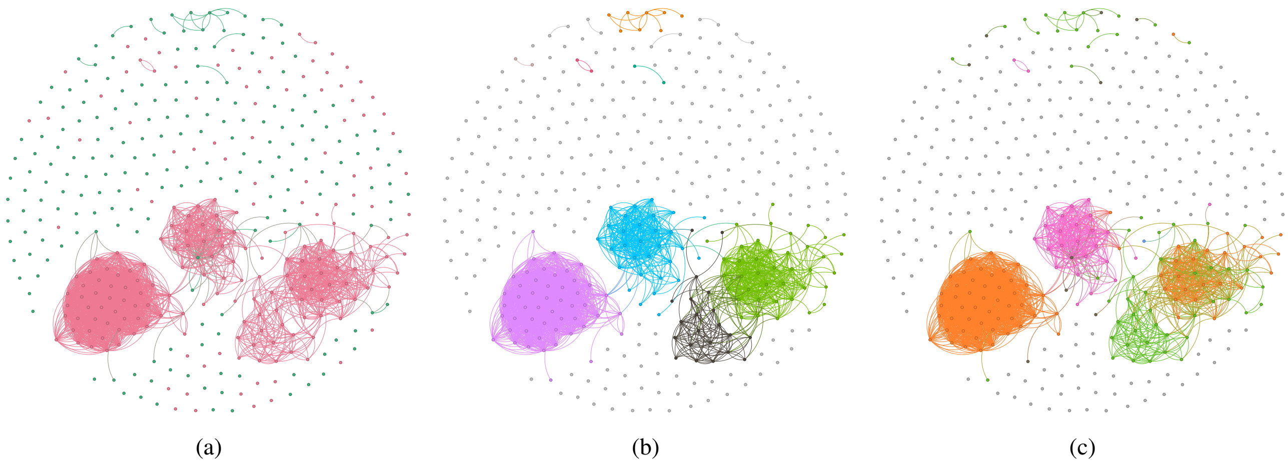

Figure 3: Graph of influences between users $N_C = 200, N_N = 200$. Each edge represents an edge identified by CCM. The edge color is simply an average color between the vertices. (a) Pink vertices are known coordinating users. Green vertices are known normal users. (b) Vertex color represents the community identified (Blondel et al. 2008). (c) Vertex color represents the topic of discussions of each user identified by CCM ● – General, ● – Trump vs. Hillary, ● – News, ● – Democratic Party, ● – Emotions.

from ${}^{N}C_2$ to $n \times {}^{N/n} C_2$. The relative decrease in run time is:

$$\frac{n \times {}^{N/n} C_2}{{}^{N}C_2} = \frac{n\frac{N/n(N/n-1)}{2}}{\frac{N(N-1)}{2}} = \frac{N-n}{n(N-1)}$$
$$\sim \frac{1}{n}, \text{ Given } N \gg n$$

This is a huge increase in performance in the best case. A suitably engineered clustering technique could achieve nearly equal clusters and hence can achieve this much performance increase in terms of computational time to our model.

We experimented with different clustering techniques, and compared those results with the sub-communities identified above as the baseline. For a comparison metric, Adjusted Rand Score (Steinley 2004) was used. Since we detected 5 sub-communities using community detection for the above sample, for comparison, we used $n = 5$ as the number of components (clusters) for each clustering method, since we observed 4 large visible clusters and a small cluster at the top of the graph in Figure 3b. Table 1 shows that NMF (Non-negative Matrix Factorization) topic modelling yields the best results out of the tested methods.

Topic modelling. NMF (Févotte and Idier 2011) is a matrix factorization technique that decomposes a non-negative $W \times H$ sized matrix into two matrices of size $W \times n$ and $n \times H$ as a product. n is a significantly smaller number than W and H. Due to the clustering property of NMF, semantically related terms are automatically grouped, forming distinct topics. In order to perform NMF, a document term matrix is constructed while TF-IDF weight adjustment is applied to the dataset to ensure term importance. Given an n, NMF decomposes this matrix into two matrices: (1) Document term matrix ($W \times n$) - Each row represents a document, and each column represents a topic, indicating the document's distribution over topics. (2) Term-topic matrix ($n \times H$) - Each row represents a topic, and each column represents a term, indicating the importance of each term within each topic. The challenge here is to find the least number of topics that partitions the dataset into semantically different subsets. Practically, maximizing the Average Silhouette Score (Rousseeuw 1987) can be recommended to determine the number of clusters n.

Table 1: Comparing different clustering techniques with the identified communities.

Method	Adjusted Rand Score
Baseline (Communities)	1
NMF	**0.38**
K-Means	0.09
DBSCAN	0.11
OPTICS	0.12
Feature Agglomeration	0.21

Observations and Optimization Methodology. Each tweet was treated as a document. Both English and Russian stop-words were removed and the documents were vectorized using TF-IDF vectorization. Then, NMF was applied to the matrix constructed by concatenating the TF-IDF vectors. For the following samples, the number of topics was chosen as 5 to run NMF due to the observations made in Figure 3b. To derive the cluster of each of user they belongs to, each tweet of a user is concatenated into a single document. Then, the trained NMF was used to predict the topic to which that

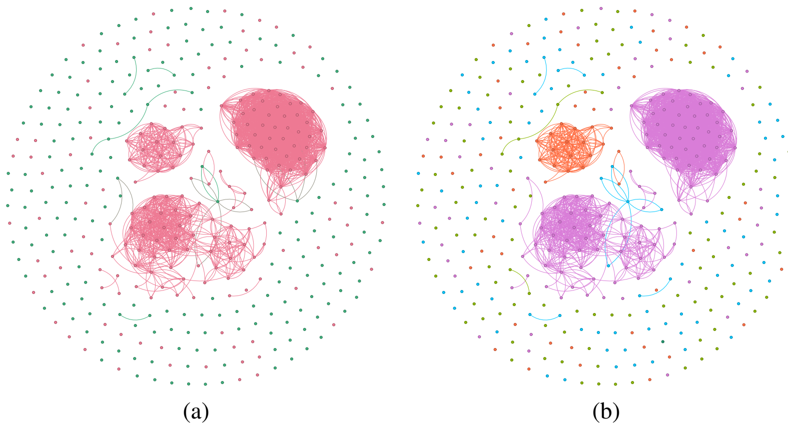

Figure 4: Graph of influences between users $N_C = 200, N_N = 200$. Each edge represents an edge identified by CCM after isolating user groups by topics. The edge color is simply an average color between the vertices. (a) Pink vertices are known coordinating users. Green vertices are known normal users. (b) Vertex color represents the topic of discussions of each user. ● – News, ● – General, ● – Democratic Party.

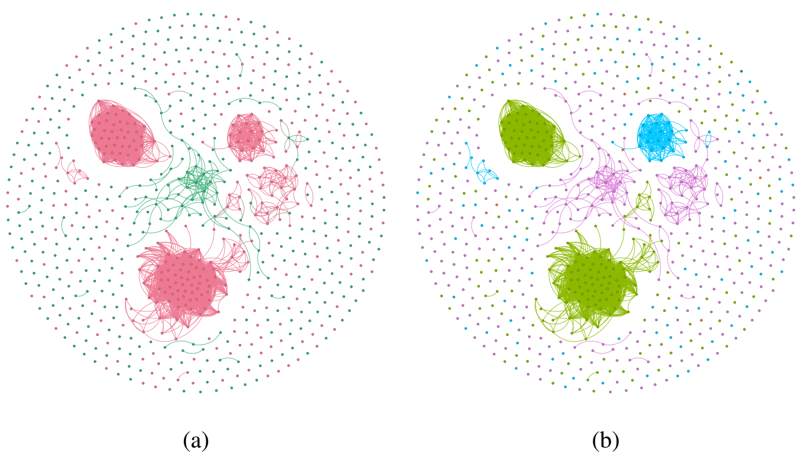

Figure 5: Graph of influences between users $N_C = 400, N_N = 400$. Each edge represents an edge identified by CCM after isolating user groups by topics. The edge color is simply an average color between the vertices. (a) Pink vertices are known coordinating users. Green vertices are known normal users. (b) Vertex color represents the topic of discussions of each user. ● – General, ● – Politics, ● – News.

long document belongs. The percentage shows the proportion of the number of people who belong to each topic out of everyone who was tested. The following are the top words that appeared in topics along with our own interpretation of the topic in a single word/phrase.

- need, make, think, life, want, know, people, just, like, don – General (44.2%)
- campaign, debate, cnn, says, vote, politics, donald, clinton, hillary, trump – Trump vs. Hillary (33.9%)
- killed, new, state, cbs, man, says, kansas, police, world, news – News (15.8%)
- far, muslim, isis, president, american, hillary, america, usa, obama, tcot – Democratic Party (5.5%)
- ll, let, heart, oh, fall, hate, song, true, life, love – Emotions (0.6%)

Figure 3c shows a graph of the users colored by the topic they are associated with. It is apparent that some topics clearly overlap with the clusters we identified using community detection in Figure 3b.

For NMF to be used as an optimization step, we cluster users using the topic. Then, we do pairwise cross mapping for each user pair inside the cluster. We evaluated the performance of our CCM model while exploring the impact of

Table 2: Comparing original results and results with topic clustering including runtimes in minutes. CC – the number of known coordinating - coordinating pairs detected by the model, CN – the number of known coordinating - normal pairs detected by the model, and NN – the number of normal - normal pairs detected by the model.

Dataset	Method	Runtime	CC	CN	NN	Precision	Recall	F1 Score
$N_C = 100, N_N = 100$	CCM	78	610	41	10	87.3%	**62.0%**	72.5%
	CCM + NMF	41	524	7	7	**91.0%**	61.0%	**73.0%**
$N_C = 200, N_N = 200$	CCM	642	2319	63	22	80.0%	**66.0%**	72.0%
	CCM + NMF	133	1818	12	7	**91.4%**	64.0%	**75.3%**
$N_C = 400, N_N = 400$	CCM	1173	4453	123	153	66.1%	**52.2%**	58.4%
	CCM + NMF	802	4106	13	119	**69.8%**	49.8%	**58.1%**

incorporating NMF on accuracy and runtime for different sizes of datasets. Table 2 summarizes the results.

CCM combined with NMF demonstrated higher precision than CCM alone, identifying a greater proportion of true coordinating pairs among those detected. Recall remained relatively consistent across both methods due to the reduction in search space, suggesting similar abilities to detect existing coordinated pairs. CCM + NMF consistently exhibited faster runtimes compared to CCM alone. This suggests that topic clustering can significantly improve efficiency without compromising accuracy. CCM + NMF results in lower precision and recall for the $N_C = 400, N_N = 400$ case. This highlights potential scalability challenges for CCM.

Figure 4 shows the derived influence graphs with this optimization for $N_C = 200$ and $N_N = 200$. Figure 5 shows results for $N_C = 400$ and $N_N = 400$.

4.5 Baseline comparisons

We choose the following baselines to compare our results.

1. *LCN + HCC (Weber and Neumann 2021).* This approach aims to identify coordinated communities using community detection on user similarity graphs. The temporal aspect is considered by a windowing mechanism. We set the window length parameter to 10 days.
2. *Tweet language.* Since most (82%) of the data in the coordinated set of users are in Russian and most (93%) of the data in the noise data are in English, we compare our results with the results of a naive model that only uses the language to determine the coordination status. This model simply classifies a user to be coordinated if the language is Russian.
3. *AMDN-HAGE.* (Sharma et al. 2021) This is the SOTA for identifying coordinated users. We use the same set of hyperparameters except the threshold to determine the output influence values. Instead, we maximize the F1 score to determine it.

CCM achieved the highest precision, indicating a superior ability to accurately identify true coordinating pairs among those detected. This suggests CCM's effectiveness in minimizing false positives, a crucial aspect of coordinated user detection. AMDN-HAGE exhibited the highest recall, suggesting its strength in detecting the majority of existing coordinating users. However, its relatively low precision indicates a higher propensity for false positives, perhaps due to the limited timeframe of the dataset. CCM + NMF achieved the highest F1 scores for datasets with 200 and 400 users, demonstrating a favorable balance between precision and recall. This highlights its potential to provide more comprehensive and accurate coordination detection compared to the other baselines. However, its performance for 800 users was lower compared to LCN + HCC, indicating potential room for further optimization.

4.6 Leader-follower behaviour

Recall that influence is a directional relationship between users. A leader on an OSN could be someone who originates content or significantly contributes to the spread of content, ideas or trends in the network. Such leaders can be identified by examining how often they are retweeted/mentioned, having high degree centrality in the influence graph. On the other hand, a follower is someone who consumes or amplifies the content of leaders. Vertices whose indegree is high but outdegree is relatively low in the influence graph could be a user with a follower personality.

Define *net-degree* to be the difference between the outdegree($\deg^+(v)$) and indegree($\deg^-(v)$) i.e., $\text{ndeg}(v) = \deg^+(v) - \deg^-(v)$. We inspected the influence graphs and checked the users who have the top net-degrees. To verify our results, for each user in the sample, we listed the number of times they were retweeted and the number of times they were mentioned. We recorded the percentile they belong in both categories. Table 4 demonstrates our results. The following are the user display names associated with the top users and some details about them (Russian names are translated to English).

- u_1: Open Russia – Open Russia constitutes a political organization established by the exiled Russian businessman Mikhail Khodorkovsky (Mikhail Khodorkovsky 2014).
- u_2: John Betts – Retired CFO, Patriot, Conservative Constitutionalist, 2nd Amd Supporter, Penn State Grad - Political Science (according to their profile description on Twitter).
- u_3: Ramzan Kadyrov – A Russian politician, currently the head of the Chechen Republic
- u_4: Moscow Bulletin – A bulletin service

Table 3: Results for detecting coordinated users using different methods. N_C - number of coordinating users in the dataset, N_N - number of normal users in the dataset.

Method	$N_C = 100, N_N = 100$			$N_C = 200, N_N = 200$			$N_C = 400, N_N = 400$		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Tweet language	64.0%	80.0%	71.1%	66.0%	81.0%	72.7%	66.0%	75.0%	70.2%
LCN + HCC	76.1%	63.0%	68.9%	77.3%	65.0%	70.6%	**81.5%**	70.4%	**75.5%**
AMDN-HAGE	50.0%	**98.0%**	66.2%	50.4%	**100%**	67.0%	50.6%	**100%**	67.2%
CCM	87.3%	62.0%	72.5%	80.0%	66.0%	72.0%	66.1%	52.2%	58.4%
CCM + NMF	**91.0%**	61.0%	**73.1%**	**91.4%**	64.0%	**75.3%**	70.1%	49.3%	57.9%

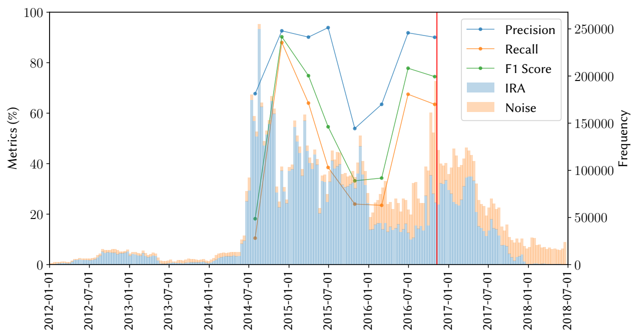

Figure 6: Results at different times for the same set of users. This indicates that the IRA Twitter attackers were performing coordinated attacks even before 2016 US elections.

- u_5: Bulletin of Novosibirsk – A bulletin service

According to the above information, the influence graph combined with net-degree sorting was able to pick important users without prior knowledge of the content they post, thus supporting the reliability and effectiveness of the CCM methodology.

4.7 Uncovering coordinated behaviours

We applied our model to the same set of users but to different periods of time. Interestingly, we get better results between November 2014 to July 2015 (See Figure 6) compared to 2016 election times, which indicates higher coordination at that time. Upon inspection, we could observe that almost all the discussions were in Russian and they are related to mostly Russian and Ukrainian politics. There existed minor discussions related to US politics as well. Our results show that the IRA has been politically influencing different parts of the world even before 2016 US Elections even though the dataset was released due to their anomalous activity in 2016. The following are the top words translated from Russian to English that appeared in the identified topics in relevant time periods.

- 4 months since November 2014
  - politics, Vladimir, news, sanctions, rf, anti-sanctions, stoptank, Putin, Russia, EU
  - will happen, prodigal, politics, Kiev tell the truth, defeat, plan, provocation of Kiev, Poroshenko, news, Ukraine
  - next, situation, difference, interesting, battle of oligarchs, happening, provocation of Kiev, Kiev tell the truth, Kievsbilboing, Ukraine
  - politics, world, read, interesting, retweet against Obama, Obama, politics, Obama, American plague, usa
  - read, ready, looks like, interesting, battle of the oligarchs, provocation of Kiev, Kievsbilboing, Kiev tell the truth, gas sector, Ukraine
- 4 months since March 2015
  - EU, battle of the oligarchs, putin, alien, worthy, rf, quot, opinion, ukraine, news
  - sanctions, rf, politics, politics, read, interesting, Obama, Ukraine, return California, USA
  - foreign ministry, Poroshenko, sanctions, politics, coming, Klimkin, not easy, negotiations, Ukraine, Russia
  - zelenskyrun, god, national, idea, Russia, Russians, read, written, interesting, Russian spirit

5 Conclusion

In this work, we proposed an approach to identify causally linked coordinating user pairs by employing convergent cross mapping of their activity traces. We consider a coordinated community as a dynamic system of variables devoid of external influences. The clustered nature of the influence graphs motivated us to pre-cluster users as a preliminary step before applying CCM, thereby reducing the overall search space. In conclusion, CCM demonstrates competitive performance in detecting coordinated users on Twitter, particularly excelling in precision. Its ability to identify causal relationships between users' activities offers a unique advantage over traditional content-based or network-based methods.

For future work, computation can be parallelized for better runtimes since pairwise comparisons of users are independent of each other. An important direction for future research could be to study how the influence graphs are evolve over time, and how the influence of important nodes in the graph is reflected on the users who follow them over the time. Future research should also explore strategies to enhance recall and scalability for large-scale applications.

Table 4: Users with top net-degree in the derived influence graph without optimizations. The number of tweets, retweets and mentions are calculated within the sample time window of 4 months from July 2016 with $N_C = 200$, $N_N = 200$. User-ids are hidden due to Twitter terms of service.

User	IRA user?	Number of tweets	Net-degree (Indegree, Outdegree)	Retweets (Percentile)	Mentions (Percentile)
u_1	Yes	788	14 (8, 22)	34 (94.8%)	36 (93.8%)
u_2	No	6729	8 (20, 28)	88 (95.8%)	112 (96.5%)
u_3	Yes	1061	7 (14, 21)	89 (96.0%)	91 (95.5%)
u_4	Yes	1257	6 (19, 25)	3 (75.1%)	5 (74.5%)
u_5	Yes	891	5 (14, 19)	6 (81.3%)	6 (77.2%)

References

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008.

Clark, A. T.; Ye, H.; Isbell, F.; Isbell, F.; Deyle, E. R.; Cowles, J.; Tilman, G. D.; Tilman, G. D.; and Sugihara, G. 2015. Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, 96 5: 1174–81.

Cobey, S.; and Baskerville, E. B. 2016. Limits to Causal Inference with State-Space Reconstruction for Infectious Disease. *PLoS ONE*, 11.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Frossard, V.; Rimet, F.; and Perga, M.-E. 2018. Causal networks reveal the dominance of bottom-up interactions in large, deep lakes. *Ecological Modelling*, 368: 136–146.

Févotte, C.; and Idier, J. 2011. Algorithms for Nonnegative Matrix Factorization with the β-Divergence. *Neural Computation*, 23(9): 2421–2456.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Hristakieva, K.; Cresci, S.; Da San Martino, G.; Conti, M.; and Nakov, P. 2022. The Spread of Propaganda by Coordinated Communities on Social Media. In *ACM WEBSCI*.

Krakovská, A.; and Hanzely, F. 2016. Testing for causality in reconstructed state spaces by an optimized mixed prediction method. *Physical review. E*, 94 5-1: 052203.

Lee, K.; Caverlee, J.; Cheng, Z.; and Sui, D. Z. 2011. Content-Driven Detections of Campaigns in Social Media. In *CIKM*.

Lee, K.; Caverlee, J.; Kamath, K. Y.; and Cheng, Z. 2012. Detecting Collective Attention Spam. In *Proceedings of WICOW/AIRWeb - WebQuality '12*.

Luo, C.; Zheng, X.; and Zeng, D. 2014. Causal Inference in Social Media Using Convergent Cross Mapping. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, 260–263.

Luo, M.; Kantz, H.; Lau, N.; Huang, W.; and Zhou, Y. 2015. Questionable dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proceedings of the National Academy of Sciences*, 112: E4638 – E4639.

Magelinski, T.; Ng, L. H. X.; and Carley, K. M. 2021. A Synchronized Action Framework for Responsible Detection of Coordination on Social Media. arxiv:2105.07454.

Mikhail Khodorkovsky. 2014. Mikhail Khodorkovsky's Initiatives - Mikhail Khodorkovsky — khodorkovsky.com. https://khodorkovsky.com/programmes/initiatives/.

Mueller, R. S.; Internet Research Agency, L.; and States, U. 2018. UNITED STATES OF AMERICA V. INTERNET RESEARCH AGENCY LLC.

Nizzoli, L.; Tardelli, S.; Avvenuti, M.; Cresci, S.; and Tesconi, M. 2021. Coordinated Behavior on Social Media in 2019 UK General Election. arxiv:2008.08370.

Pacheco, D.; Hui, P.-M.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. *Proceedings of AAAI ICWSM*, 15.

Permanent Select Committee on Intelligence. 2018. Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements. https://intelligence.house.gov/social-media-content/default.aspx.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Sharma, K.; Zhang, Y.; Ferrara, E.; and Liu, Y. 2021. Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours. In *KDD*.

Shchur, O.; Biloš, M.; and Günnemann, S. 2020. Intensity-Free Learning of Temporal Point Processes. arxiv:1909.12127.

Steinley, D. 2004. Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods*, 9(3): 386–396.

Sugihara, G.; May, R.; Ye, H.; Hsieh, C.-h.; Deyle, E.; Fogarty, M.; and Munch, S. 2012. Detecting Causality in Complex Ecosystems. *Science*, 338(6106): 496–500.

Takens, F. 1981. Detecting strange attractors in turbulence. In Rand, D.; and Young, L.-S., eds., *Dynamical Systems and Turbulence, Warwick 1980*, 366–381. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-38945-3.

Tsonis, A. A.; Deyle, E. R.; Ye, H.; and Sugihara, G. 2018. *Convergent Cross Mapping: Theory and an Example*, 587–600. Cham: Springer International Publishing. ISBN 978-3-319-58895-7.

van Nes, E. H.; Scheffer, M.; Brovkin, V. A.; Lenton, T. M.; Ye, H.; Deyle, E. R.; and Sugihara, G. 2015. Causal feedbacks in climate change. *Nature Climate Change*, 5: 445–448.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. *NeurIPS*, 30.

Weber, D.; and Falzon, L. 2022. Temporal Nuances of Coordination Network Semantics. arxiv:2107.02588.

Weber, D.; and Neumann, F. 2021. Amplifying Influence through Coordinated Behaviour in Social Networks. *SNAM*, 11(1).

Ye, H.; Deyle, E. R.; Gilarranz, L. J.; and Sugihara, G. 2015. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports*, 5.

Zhang, N.; Wang, G.; and Tsonis, A. A. 2018. Dynamical evidence for causality between Northern Hemisphere annular mode and winter surface air temperature over Northeast Asia. *Climate Dynamics*, 52: 3175–3182.

Zhang, Y.; Sharma, K.; and Liu, Y. 2021. VigDet: Knowledge Informed Neural Temporal Point Process for Coordination Detection on Social Media. arxiv:2110.15454.

6 Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this research tries to promote autonomy of users on social media by detecting mass campaigns.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the content of the paper is outlined in the abstract.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, the results are verified in Sections 4.5, 4.6, and 4.7.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, they are discussed in Sections 4.1 and 4.5.
   (e) Did you describe the limitations of your work? Yes, the limitations in runtime, recall, and scalability are discussed at the end of Section 4.4 and Section 4.5.
   (f) Did you discuss any potential negative societal impacts of your work? No, because we do not see a potential negative impact of our work on the society.
   (g) Did you discuss any potential misuse of your work? No, because we do not see a potential of such misuse of our work.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings?
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we did.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, the instructions are given in the article.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, the experimental setup is provided in Section 4.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, because the core machine learning component (KNN classifier) is a deterministic algorithm.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, that is provided in Section 4.3.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? No, the evaluation method is a core component in the CCM methodology. We do not see the importance of repeating the justification to it in this article.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we discuss the improvement in precision. Our work significantly increases precision in identifying coordinated users compared to

baselines. Therefore, reducing the cost of misclassifications is a key contribution in our work.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? No, because we use common Python libraries such as Numpy, Scikit-Learn, and Pandas.
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, because we comply with Twitter Terms of Service.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, we do not expose details that can uniquely identify a user on Twitter.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Not applicable since we do not release our dataset. That would be a violation of Twitter's Terms of Service.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA