International Research Journal of Education and Technology
ISSN 2581-7795

# A Mechanism for Controlling Viral Fake News: A concern for Smarter City

Prof. Megha p. Nanhe

Ms. Vishakha Thakre

Dr. Ambedkar Institute of Management Studies and Research,

Deekshabhoomi, Nagpur

## ABSTRACT:

Today's social media platforms enable to spread both authentic and fake news very quickly. Some approaches have been proposed to automatically detect such "fake" news based on their content, but it is difficult to agree on universal criteria of authenticity (which can be bypassed by adversaries once known). Besides, it is obviously impossible to have each news item checked by a human. In this paper, we a mechanism to limit the spread of fake news which is not based on content. It can be implemented as a plug-in on a social media platform. The principle is as follows: a team of fact checkers reviews a small number of news items (the most popular ones), which enables to have an estimation of each user's inclination to share fake news items. Then, using a Bayesian approach, we estimate the trustworthiness of future news items, and treat accordingly those of them that pass a certain "untrustworthiness" threshold. This paper is concern form making smarter city.

**Keywords:**

Probability of fake news, Detection of fake news, Credulix Operation

## I. INTRODUCTION

The expression "fake news" has become very popular after the 2016 presidential election in the United States. Both political sides accused each other of spreading false information on social media, in order to influence public opinion. Fake news have also been involved in Brexit and seem to have played a crucial role in the French election. The phenomenon is considered by many as a threat to democracy, since the proportion of people getting their news from social media is significantly increasing.

Facebook and Google took a first concrete measure by removing advertising money from websites sharing a significant number of fake news. This, however, does not apply to websites that do not rely on such money: popular blogs, non-professional streamingchannels, or media relying on donations, to name a few. Facebook also considered labeling some news items as “disputed” when independent human fact-checkers contest their reliability. However, there cannot be enough certified human fact-checkers for a world-wide social network. While it is very easy to share fake news, it may take very long to check them, clearly too long to prevent them from getting viral.

We present Credulix, the first content-agnostic system to pre-vent fake news from getting viral. From a software perspective, Credulix is a plugin to a social media platform. From a more ab-stract perspective, it can also be viewed as a vaccine for the social network. Assuming the system has been exposed to some (small) amount of fake news in the past, Credulix enables it to prevent future fake news from becoming viral. It is important to note that our approach does not exclude other (e.g. content-based) approaches, but complements them.

At the heart of our approach lies a simple but powerful Bayesian result we prove in this paper, estimating the credibility of news items based on which users shared them and how these users treated fake news in the past. News items considered fake with a sufficiently high probability can then be prevented from further dissemination, i.e., from becoming viral.

Turning the theory behind Credulix into a system deployable in practice is a non-trivial task. In this paper we address these challenges as well. In particular, we present a practical approach to computing news item credibility in a fast, incremental manner.

The paper is organized as follows. Section 2 presents the theoretical principles behind Credulix. Section 3 presents the design and implementation of Credulix. Section 4 reports on our evaluation results. Section 5 discusses the limitations and tradeoffs posed by Credulix. Section 6 discusses related work and Section 7 concludes.

## II. Probability of fake news

Basic Fake News Detection

User Behavior. We model the behavior of a user u using the two following probabilities:

- $P_T$ (u): probability that u shares a news item if it is true.
- $P_F$ (u): probability that u shares a news item if it is fake.

The probabilities $P_T$ (u) and $P_F$ (u) are assumed to be independent between users. In practice, this is the case if the decision to share a news item X is mainly determined by X itself.

We obtain estimates of $P_T$ (u ) and $P_F$ (u) for each user based on the user’s behavior (share / not share) with respect to fact-checked items. For any given user u, let $v_T$ (u) (resp. $s_T$ (u)) de-note the number of fact-checked true news items viewed (resp. shared) by u, and $v_F$ (u) (resp. $s_F$ (u)) the number of fact-checked fake news items viewed (resp. shared) by u. We call the tuple ($v_T$ (u), $s_T$ (u), $v_F$ (u), $s_F$ (u)) the User Credulity Record (UCR) of u.

User behaviour has been modeled similarly in prior work, for instance, in Curb. In Curb, users exposed to a news item decide probabilistically whether to share it, potentially exposing all their followers to it. Curb relies on very similar metrics, namely the numbers of viewed and shared items for each user, and the probabilities that users would share true or false news items.

For any given user u, we define the following functions, based on the UCR:

- $\beta_1(u)=(s_T(u)+1)/(v_T(u)+2)$
- $\beta_2(u)=(s_F(u)+1)/(v_F(u)+2)$

- $\beta_3(u)=(v_T(u) - s_T(u)+1)/(v_T(u)+2)$
- $\beta_4(u)=(v_F(u) - s_F(u)+1)/(v_F(u)+2)$

According to Laplace's Rule of Succession, we have $P_T$ (u) = $\beta_1$(u)and$P_F$(u)= $\beta_2$(u). Probability of a News Item Being Fake.

## III. Fast Fake News Detection

Credulix' measure of credibility of a news item X is the probability p(V , S) that X is fake. An obvious way to compute this probability is to recalculate p(V , S ) using eq. (1) each time X is viewed or shared by a user. Doing so, however, would be very expensive in terms of computation. Below, we show an efficient method for computing news item credibility. We first describe the computation of UCRs, and then present our fast, incremental approach for computing news item credibility using item ratings and UCR scores. This is crucial for efficiently running Credulix in practice.

Computing User Credulity Records (UCRs). Recall that the four values ($v_T$ (u), $s_T$ (u), $v_F$ (u), $s_F$ (u)) constituting a UCR only concern fact-checked news items. We thus update the UCR of user u (increment one of these four values) in the following two scenarios.

(1) When u views or shares a news item that has been fact-checked (i.e., is known to be true or fake).

(2) Upon fact-checking a news item that u had been exposed to.

In general, the more fact-checked news items a user u has seen and shared, the more meaningful u's UCR. Users who have not been exposed to any fact-checked items cannot contribute to Credulix.

## IV. CREDULIX AS A SOCIAL MEDIA PLUGIN

Credulix can be seen as a plugin to an existing social network, like, for instance, Facebook's translation feature. The translator observes the content displayed to users, translating it from one language to another. Similarly, Credulix observes news items about to be displayed to users and tags or suppresses those considered fake.

Selective Item Tracking. Every second, approximately 6000 new tweets appear on Twitter and 50000 new posts are created on Facebook. Monitoring the credibility of all these items would pose significant resource overhead. With Credulix, each view / share event requires an additional update to the news item's meta-data. However, we do not need to keep track of all the items in the system, but just the ones that show a potential of becoming viral.

Credulix requires each item's metadata to contain an additional bit indicating whether that item is tracked. The rating of item X is only computed and kept up to date by Credulix if X is tracked.

Interaction with the Social Media Platform. We consider two basic operations a user u can perform:

- Sharing a news item and
- Viewing her own news feed.

Sharing is the operation of disseminating a news item to all of u's followers (e.g., tweeting, sharing, updating Facebook status etc.). Viewing is the action of refreshing the news feed, to see new posts shared by users that u follows. In the following, we describe how these operations are performed in a social network platform (in-spired by Twitter) without Credulix (Baseline) and with Credulix. We assume that, like in Twitter, all users' news feeds are represented as lists of item IDs and stored in memory, while the item contents are stored in an item data store .

Viewing News Feed with Credulix.Credulix augments the View operation in two ways. First, after the news feed articles are retrieved from the data store,Credulix checks the ratings of the items, filtering out the items with a high probability of being fake. Second, if u's news feed contains tracked items, Credulix updates the rating of those items using u's UCR view score. Hence, a supplementary write to the data store is necessary, compared to the Baseline version, for storing the items' updated ratings. Again, we do this in the background, not impacting user request latency.

## V. EVALUATION

In this section, we evaluate our implementation of Credulix as a stand-alone Java plugin. We implement a Twitter clone where the share and view operation executions are depicted. We refer to the Twitter clone as Baseline and we compare it to the variant with Credulix plugged in, which we call Credulix. For the data store of the Baseline we use Twissandra's data store,running Cassandra version 2.2.9.

The goals of our evaluation are the following. First, we explore Credulix's fake news detecting efficiency. Second, we measure the performance overhead of our implementation. More precisely, we show that:

(1) Credulix efficiently stops the majority of fake news from be-coming viral, with no false positives. Credulix reduces the number of times a viral fake news item is viewed from hun-dreds of millions to hundreds of thousands (in Section 4.2).

(2) Credulix succeeds in stopping the majority of fake news from becoming viral for various user behaviors in terms of how likely users are to share news items they are exposed to (in Section 4.3).

(3) Credulix's impact on system performance is negligible for both throughput and latency (in Section 4.4).

## VI. Fake News Detection Relative to Sharing Probability

In Figure 6 we plot the percentage of fake items displayed with Credulix for two graph sizes. On a smaller graph of 1M users generated with the SNAP generator, Credulix achieves a lower fake item detection rate. This is because the impact of fact-checked items is smaller on a small graph, leading to fewer users with relevant UCR scores. This result suggests that on a real social graph that is larger than the one we use, Credulix would be more efficient than in our experiments.

## VII. CREDULIX Overhead

In this experiment, we evaluate Credulix' impact on user operations' (viewing and sharing) throughput and latency. We present our results for four workloads, each corresponding to a value of msp discussed above. We present results for two social graph sizes: 41M users, and 1M users, with 16 worker threads serving. The impact of garbage collection is stronger with Credulix than with Baseline, as Credulix creates more short-lived objects in mem-ory to orchestrate its background tasks. In addition to the intrinsic differences between users discussed above, garbage collection also significantly contributes to the high standard deviation observed in all latencies.

## VIII. DISCUSSION AND LIMITATIONS

We believe that Credulix is a good step towards addressing the fake news problem, but we do not claim it to be the ultimate solution. Credulix is one of many possible layers of protection against fake news and can be used independently of other mechanisms. With Credulix in place, news items can still be analyzed based on content using other algorithms. It is the combination of several approaches that can create a strong defence against the fake news phenomenon. This section discusses the limitations of Credulix.

News Propagation.Credulix does not prevent users from actively pulling any (including fake) news stories directly from their sources. Credulix identifies fake news on a social media platform and, if used as we suggest, prevents users from being notified about other users sharing fake news items.

Manual Fact-Checking.Credulix relies on manual fact-checking and thus can only be as good as the fact-checkers. Only users who have been exposed to manually fact-checked items can be leveraged by Credulix. However, fact-checking a small number of popular news items are sufficient to obtain enough users with usable UCRs. Fact-checking a small number of news items are feasible, especially given the recent upsurge of fact-checking initiatives.

User Behavior.Credulix's algorithm is based on the assumption that among those users exposed to fact-checked news items, some share more fake items than others. Analogous assumptions are commonly used in other contexts such as recommender systems. For example, a user-based recommender system would not be useful if all users behaved the same way, i.e. everybody giving the same ratings to the same items. Note that, even in the case where all users did behave the same, running Credulix would not negatively impact the behavior of the system: Credulix would not detect any fake news items, nor would it classify fake news items as true. Our approach shines when the inclination of most users to share fake news does not change too quickly over time. Looking at social media today, some people seem indeed to consciously and consistently spread more fake news than others. In fact, many systems that are being successfully applied in practice (e.g. reputation systems, or systems based on collaborative filtering) fundamentally rely on this same assumption.

Filtering News. One could argue that removing some news items from users' news feeds might be seen as a limitation, even as a form of censorship. But social media already take that liberty as they display to users only about 10% of the news that they could show. Rather than censorship, Credulix should be viewed as an effort to ensure the highest possible quality of the items displayed, considering the credibility of an item to be one of the quality criteria.

## IX. CONCLUSION

We presented Credulix, the first content-agnostic system to detect and limit the spread of fake news on social networks with a very small performance overhead. Using a Bayesian approach, Credulix learns from human fact-checking to compute the probability of falsehood of news items, based on who shared them. Applied to a real-world social network of 41M users, Credulix prevents the vast majority (over 99%) of fake news items from becoming viral, with no false positives.

## References:

**https://arxiv.org/pdf/1808.09922.pdf**