# FActCheck: Keeping Activation of Fake News at Check

Extended Abstract

Ajitesh Srivastava
University of Southern California
Los Angeles, California
ajiteshs@usc.edu

Rajgopal Kannan
US Army Research Lab-West
Los Angeles, California
rajgopal.kannan.civ@mail.mil

Charalampos Chelmis
University at Albany - SUNY
Albany, New York
cchelmis@albany.edu

Viktor K. Prasanna
University of Southern California
Los Angeles, California
prasanna@usc.edu

## ABSTRACT

The diffusion of fake news has become a crucial problem in recent years. One way to battle it is to propagate the corresponding real news. To achieve this goal, we find a set of individuals who are likely to receive the fake news so that they can test its credibility, and when they propagate the corresponding real news, it is likely to reach a large number of individuals. For this problem, we propose a polynomial time greedy algorithm (AFC) which provides $(1 - 1/e - \epsilon)$-approximation. We further optimize the runtime of AFC by developing a fast graph-pruning heuristic (RAFC) that performs as well as AFC in checking the spread of fake news. Our experiments on real-world networks demonstrate that our approach outperforms popular methods in social network analysis literature.

## KEYWORDS

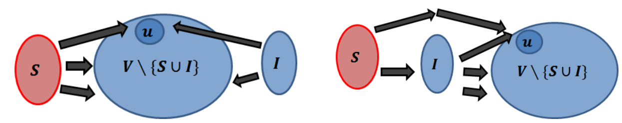
Information Diffusion; Social Networks; Fake News

**ACM Reference Format:**
Ajitesh Srivastava, Rajgopal Kannan, Charalampos Chelmis, and Viktor K. Prasanna. 2018. FActCheck: Keeping Activation of Fake News at Check. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018*, IFAAMAS, 3 pages.

## 1 INTRODUCTION

Online Social Networks have become prime sources of sharing news. Often times "fake news", a made-up story propagates into the network and becomes accepted as "news". Spreading of misinformation poses a major challenge to the society as it may influence people's opinions and cause panic. A possible remedy is to propagate the corresponding real news in the network [2, 4]. Several works [2, 4] take the *competing cascades* approach, where they attempt to ensure that an individual who receives an opinion from a given set $S$ is also likely to receive the alternate opinion from the selected set $I$. It is implicitly assumed that nodes in $I$ are already aware of the alternate opinion to be propagated. Note that in the case of containment of fake news, the alternate opinion (real news) cannot be determined unless the individuals in $I$ are also aware of what the fake news to be countered is. For instance, suppose person $A$ is propagating the news saying "NASA predicts an asteroid hitting Earth in 48 hours". When person $B$ receives this, only then can she counter it with the news that NASA has released no such statement. The competing cascades approach is thus not very effective in this problem setting as it does not enforce that $I$ must receive news from $S$ first (See Figures 1(a) and 1(b)).

*Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.* 




(a) Competing Cascades: Two sets attempting to influence a node

(b) FActCheck: A subset of those aware of the fake news can influence a node

**Figure 1: Fundamental distinction between FActCheck and Competing Cascades. FActCheck enforces the constraint that the news must pass through the set $I$.**

To address this limitation, we propose a fundamental departure from the competing cascade approach by proposing the following problem: Given a set $S$ of fake news initiators in a network, we wish to find a set of nodes $I$ such that a) the fake news from $S$ is likely to reach $I$, and b) many other nodes are reachable from $I$ (See Figure 1(b)). We refer to this problem of checking the activation of fake news as FActCheck. We propose a polynomial time algorithm called Approximate FactCheck (AFC) with $(1 - 1/e - \epsilon)$-approximation guarantee (Section 2). We also propose a heuristic called Reduced Approximate FActCheck (RAFC) that provides a quality similar to AFC while reducing the runtime significantly

***Problem Definition.*** Let $X \xrightarrow{Z} Y$ be the event of news flowing from $X$ to $Y$ through a node in $Z$ under some diffusion model. Let $S$ be the set of fake news initiators. We wish to maximize $\sigma(S, I) = \mathbb{E}\left(\sum_{u \in V \setminus S} \mathbb{I}(S \xrightarrow{I} u)\right)$, where $\mathbb{I}$ is the indicator function. We define this problem as Fake news Activation Checking (FActCheck).

*Definition 1.1 (FActCheck).* Given a graph $G(V, E)$, a seed-set $S$, a model of diffusion $\mathcal{M}$ and an integer $k$, find $I \subset V \setminus S$, $|I| = k$, that maximizes $\sigma(S, I)$.

It can be shown that this is equivalent to maximizing $P(S \xrightarrow{I} t)$, where $t$ is a randomly selected node. Henceforth, we proceed with Independent Cascade Model (ICM) as the model of diffusion due to its popularity since its introduction [3].

## 2 PROPOSED ALGORITHM FOR *FActCheck*

Our algorithm is based on generation of "Pruned Reverse Reachable" sets (PRR) which is the set of all nodes that connect at least one node in $S$ to a randomly selected node, in one instance of live-edge graph (the graph obtained by keeping an edge $(i,j)$ with probability $p(i,j)$). By construction, if $\exists v \in I$ which is present in the Pruned RR, then $\mathbb{I}(S \xrightarrow{I} t) = 1$. Therefore, if $A$ is a randomly generated Pruned RR, then $P(S \xrightarrow{I} t) = \mathbb{E}(\mathbb{I}(S \xrightarrow{I} t)) = \mathbb{E}(\mathbb{I}(|A \cap I| > 0))$.

This probability can be estimated by generating $\theta$ number of Pruned RRs, where $\theta$ is "very large". Once, $\theta$ Pruned RRs have been generated, we can apply greedy selection of nodes that result in largest marginal gain, to construct the desired set. The number of Pruned RRs $\theta$ that need to be generated is the same as the number of times set $A$ needs to be sampled so that greedy selection of $I$ guarantees a $(1-1/e-\epsilon)$-approximation for the optimal value of $\mathbb{E}(\mathbb{I}(|A \cap I|))$. This is given by $\theta = \frac{n(2+2\sqrt{2}\epsilon/3)(\log\binom{n}{k}+\log n+\log\log_2 n)}{2\epsilon^2 OPT}$ [6], where $n = |V \setminus S|$ and $OPT$ is the optimal value of $nP(S \xrightarrow{I} t)$. We refer to this method of generating $\theta$ PRRs and applying greedy algorithm to solve FActCheck as Approximate FActCheck (AFC).

***Graph Reduction.*** One drawback of using AFC for FActCheck is that due to small size of $OPT$ and large graph size, the number of PRRs required may be very high as $\theta \propto n/OPT$. To address this, we propose to first reduce the size of the graph before applying AFC to obtain a reasonable solution in less time.

Theorem 2.1. *Let $V_S$ be the set of nodes that are reachable from $S$ in at least $r$ instances of a live-edge graph in $R = \alpha \log n/\mu$ randomly generated instances for some $\alpha$. Assuming $r \leq \log n$, with probability $1-\delta$, $P(\cup_{v \in V \setminus V_S}(S \rightarrow v)) < \mu$, where $\delta = \frac{\alpha^r}{(r-1)!(1-\mu)^r}\frac{(\log n)^r}{n^{\alpha-1}}$.*

We simulate the diffusion process starting with the source $S$, $\alpha \log n/\mu$ times. All nodes that are reached in at least $r$ simulations are added to $V_S$. Then, we run AFC on $G_S$, the graph induced by $V_S$ (where $|V_S \setminus S| = n_S \leq n$) which is a smaller than $V$. We refer to this algorithm as Reduced AFC (RAFC).

## 3 EXPERIMENTS

***Setup.*** The probability of influence for edge $(u,v)$ was set to $p(u,v) = 1/d_v$, where $d_v = \sum_j weight(v,j)$ is the sum of the outgoing edge weights. The seedset $S$ was set to the 50 nodes with highest out-degree ($\sum_v p(u,v)$). Datasets are summarized in Table 1.

***Quality of Reduction.*** We performed a series of experiments to measure the effect of reduction on the execution time and quality of the results obtained, i.e., $\sigma(S,I)$. For RAFC, we used $\mu = 0.1, 0.01, 0.001, \ldots$ to reduce the graph size. For LJ, no significant reduction in size was obtained for $\mu = 0.01$ and higher and therefore, only the result for $\mu = 0.1$ has been reported. Table 1 shows the comparison of reduction in $\sigma(S,I)$ vs speedup obtained by RAFC compared to AFC. HEPT being a small dataset did not show a significant speedup. Maximum speedup was seen in Twitter dataset (102x) without much compromise in quality.

***Baselines.*** (i) Outdegree ($\sum_{j \in V \setminus S} p(v,j)$), (ii) PageRank [5], (iii) PPR (Personalized PageRank) [5] where random walk starts from a random node in $S$, (iv) FSBC (Fixed Source Betweenness Centrality) - where we only count shortest path with source in $S$ and edge weights are $-\log(p(i,j))$, (v) SmartDegree - calculated for node $v$ as $\sum_{u \in S} p(u,v) \sum_{j \in V \setminus S} p(v,j)$.

**Table 1: Datasets and details of graph reduction obtained**

| Dataset | | HEPT[1] | Twitter[1] | LJ[2] |
|---|---|---|---|---|
| # of nodes | | 15,233 | 3,919,215 | 4,847,571 |
| # of edges | | 62,774 | 5,399,949 | 68,475,391 |
| Quality reduction, speedup | $\mu = 0.1$ | 2.54%, 1.81x | 3.8%, 102x | - |
| | $\mu = 0.01$ | 2.37%, 1.37x | 0.07%, 17.2x | - |
| | $\mu = 0.001$ | 0.75%, 0.48x | -0.4%, 2.30x | -0.03%, 2.45x |

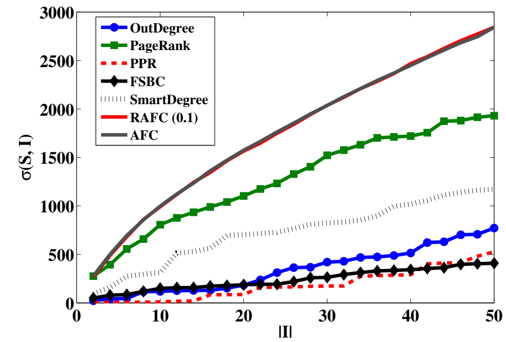



**Figure 2: Comparison of $\sigma(S,I)$ vs $|I|$ for different methods when the source is known.**

***Comparison with Baselines.*** We computed $\sigma(S,I)$ obtained using these methods along with our methods by running 10,000 simulations. The size of $I$ was varied from 1 to 50. Figure 2 shows $\sigma(S,I)$ achieved by all the methods on LJ. Results on other datasets have been omitted due to brevity, as they produced similar trends. Performance of RAFC was always found to be close to AFC. These methods outperformed the baselines by significant margins.

## 4 CONCLUSIONS

We have proposed Fake news Activation Checking (FActCheck) problem to address the challenge posed by fake news propagation in online social networks. Under Independent Cascade Model, we have given a polynomial time algorithm (AFC) with $(1-1/e-\epsilon)$-approximation guarantee. Since the runtime of AFC increases with the size of the graph, we have developed a heuristic (RAFC) that reduces the size of the graph by removing nodes that are likely to have low probability of activation, before applying AFC. Experiments have demonstrated that RAFC produces similar quality to AFC, while providing significant speed-up in runtime. Our methods were compared against popular centrality measures from social network literature. Both AFC and RAFC outperform the baselines by a large margin on several real-life networks.

## 5 ACKNOWLEDGMENTS

This work is supported by U.S. National Science Foundation under EAGER Award No.:1637372.

[1] Available at http://trec.nist.gov/data/tweets/
[2] Available at https://snap.stanford.edu/data/soc-LiveJournal1.html

## REFERENCES


[1] Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 199–208.

[2] Xinran He, Guojie Song, Wei Chen, and Qingye Jiang. 2012. Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model.. In *SDM*. SIAM, 463–474.

[3] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 137–146.

[4] Nam P Nguyen, Guanhua Yan, My T Thai, and Stephan Eidenbenz. 2012. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 213–222.

[5] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: bringing order to the web. (1999).

[6] Youze Tang, Yanchen Shi, and Xiaokui Xiao. 2015. Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 1539–1554.