The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News

Nguyen Vo and Kyumin Lee
Computer Science Department, Worcester Polytechnic Institute
Worcester, Massachusetts 01609, USA
{nkvo,kmlee}@wpi.edu

ABSTRACT

A large body of research work and efforts have been focused on detecting fake news and building online fact-check systems in order to debunk fake news as soon as possible. Despite the existence of these systems, fake news is still wildly shared by online users. It indicates that these systems may not be fully utilized. After detecting fake news, what is the next step to stop people from sharing it? How can we improve the utilization of these fact-check systems? To fill this gap, in this paper, we (i) collect and analyze online users called *guardians*, who correct misinformation and fake news in online discussions by referring fact-checking URLs; and (ii) propose a novel fact-checking URL recommendation model to encourage the guardians to engage more in fact-checking activities. We found that the guardians usually took less than one day to reply to claims in online conversations and took another day to spread verified information to hundreds of millions of followers. Our proposed recommendation model outperformed four state-of-the-art models by 11%~33%. Our source code and dataset are available at http://web.cs.wpi.edu/~kmlee/data/gau.html.

ACM Reference Format:
Nguyen Vo and Kyumin Lee . 2018. The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210037

1 INTRODUCTION

Fake news, misinformation, rumor or hoaxes are one of the most concerning problems due to their popularity and negative effects on society. Particularly, social networking sites (e.g., Twitter and Facebook) have become a medium to disseminate fake news. Therefore, companies and government agencies have paid attention to solving fake news. For example, Facebook has a plan to combat fake news[1] and the FBI has investigated disinformation spread by Russia and other countries[2].

[1]http://fortune.com/2017/10/05/facebook-test-more-info-button-fake-news/
[2]http://bit.ly/FBIRussian

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210037

Figure 1: An example of fact-checking activity.

To verify correctness of information, researchers proposed to (i) employ experts, who can fact-check information [59], (ii) use systems that can automatically check credibility of news [19, 33, 46]; and build models to detect fake news [7, 24, 35, 42, 53]. In 2016, Reporter Lab reported that the number of fact-checking websites went up by 50%[3]. However, fake news is still wildly disseminated on social media even when it has been debunked [36, 58].

A recent report [25] showed that 86% of American adults do not fact-check articles they read. A possible explanation for this is that people may trust content shared from their friends rather than other sources [25] or they may not have time to fact-check articles they read, or simply they may not know the existence of these fact-check websites. It means that merely debunking fake news is not enough, and these systems are not fully utilized.

Furthermore, it has been shown that once absorbing misinformation from fake news, individuals are less likely to change their beliefs even when the fake news are debunked. If the idea in the original fake news is especially similar to individuals' viewpoints, it will be even harder to change their minds [12, 40]. Therefore, it is needed to deliver verified information quickly to online users before fake news reaches them. To achieve this aim, the volume of verified content should be large enough on social networks, so that online users may have a higher chance to be exposed to legitimate information before consuming fake news from other sources.

In this paper, we propose a framework to further utilize fact-checked content. Particularly, we collect a group of people and stimulate them to disseminate fact-checked content to other users. However, achieving the goal is challenging because we have to solve the two following problems: (**P1**) How can we find a group of people (e.g. online users) who are willing to spread verified news? (**P2**) How can we stimulate them to disseminate fact-checked news/information?

[3]http://reporterslab.org/global-fact-checking-up-50-percent

To deal with the first problem **(P1)**, we may deploy bots [27, 49] to disseminate information but it may violate terms of services of online platforms due to abusing behavior. Another approach is to hire crowd workers [29] and cyber troops to shape public opinion [5]. However, this approach may cost a lot of money and is difficult to deploy in larger scale due to monetary constraints. Inspired by [18], we propose to rely on online users called *guardians*, who show interests in correcting false claims and fake news in online discussions by embedding fact-checking URLs. Figure 1 illustrates who a guardian is and helps us to describe terminologies that we use in this paper. In the figure, two Twitter users have a conversation, in which a user @sir_mycroft accused the Clinton foundation of accepting money from *Uranium One* company in exchange for the approval of the deal between *Uranium One* and Russian government in 2009. After just 15 minutes, this false accusation was debunked by a user @Politics_PR, who referred to FactCheck.org and Snopes.com URLs as evidences to support his factual correction. We call such direct replies, which contain fact-checking URLs, *direct fact-checking tweets* (**D-tweets**). Users, who posted D-tweets, are called *direct guardians* (**D-guardians**). The user, to whom the D-guardian replied (i.e. @sir_mycroft), is called an *original poster*. In addition, we observed that @Politics_PR's response was retweeted 15 times. We call these retweeters *secondary guardians* (**S-guardians**), regardless of whether they added a comment or not inside the retweet. Their shares are called *secondary tweets* (**S-tweets**). Both *D-guardians* and *S-guardians* are called *guardians*, and both *D-tweets* and *S-tweets* are called *fact-checking tweets*. In Section 4, we investigate whether both D-guardians and S-guardians play an important role in correcting claims and spreading fact-checked information.

To cope with the second problem **(P2)**, we may directly ask the guardians to spread verified news like [28], but their response rate may be low because each guardian may be interested in different topics, and eventually, we may send unwanted requests to some of the guardians. Thus, we tackle the second problem by proposing a fact-checking URL recommendation model. By providing personalized recommendations, we may stimulate guardians' engagement in fact-checking activities toward spreading credible information to many other users and reducing the negative effects of fake news.

By addressing these two problems, we collect a large number of reliable guardians and propose a fact-checking URL recommendation model which exploits recent success in embedding techniques [32] and utilizes auxiliary data to personalize fact-checking URLs for the guardians. Our main contributions are as follows:

- We are the first work to utilize guardians, who can help spread credible information and recommend fact-checking URLs to the guardians as a pro-active way to combat fake news.
- We thoroughly analyze who guardians are, their temporal behavior, and topical interests.
- We propose a novel URL recommendation model, which exploits fact-checking URLs' content (i.e., linked fact-checking pages), social network structure, and recent tweets' content.
- We evaluate our proposed model against four state-of-the-art recommendation algorithms. Experimental results show that our model outperforms the competing models by 11%~33%.

2 RELATED WORK

In this section, we first summarize related work about fake news, rumors and misinformation. Then, we cover the prior work on URL recommendation on social network.

2.1 Fake News, Rumors and Misinformation

Although fake news on social media has been extensively studied, it still attracts the attention of communities due to its negative impact on society such as fake Russian Facebook ads and political events [4]. The majority of studies focused on classifying rumors to either true or false by exploiting different feature sets [7, 24, 35, 42, 53] or by building deep learning models [34, 45]. In natural disasters and emergency situations, misinformation was investigated as well [16, 20, 58]. Several works attempted to detect rumors as soon as possible using disputed signals [33, 58], leveraging network embedding [54] and employing collective data sources [21, 41]. However, there is no work about combating fake news once it has been debunked.

Another direction is to detect or classify stances of users (e.g. supporting or denying) toward rumors [13, 42] and to analyze how users' stances have changed over time [31, 36, 59]. In addition to studying rumors' content, researchers [23, 31] also analyzed who were involved in spreading those rumors. Since fake news can be viewed as misinformation, work about detecting content polluters [27], social bots [48] and malicious campaigns [49] are also related to our work. The following two works [14, 18] are perhaps the most closely related to our work. In particular, Hannak et al. [18] analyzed the social relationship between the fact-checking user and the fact-checked user in online conversations. [14] employed fact-checking URLs in Snopes.com as a way to understand how rumors were spread on Facebook. Our work differs from the prior works [14, 18] since we focus on guardians, their temporal behavior and topical interests, and propose a fact-checking URL recommendation model to personalize relevant fact-checking URLs.

2.2 URL Recommendation on Social Media

Chen et al., [8] proposed a content-based method to recommend URLs on Twitter. [1] proposed hashtag-based, topic-based and entity-based methods to build user profiles for news personalization. By enriching user profiles with external data sources [2, 3], Abel et al., improved URL recommendation results. Taking a similar content-based approach, Yamaguchi et al., [56] employed Twitter lists to recommend fresh URLs and [17] tried to recommend URLs on streaming data. [10] proposed an SVM based approach to recommend URLs. Dong et al., [11] exploited Twitter data to discover fresh websites for a web search engine. However, to the best of our knowledge there is no prior work employing matrix factorization models and auxiliary information to recommend URLs to guardians on Twitter. In addition to recommending URLs, researchers also focused on personalizing who to follow [6], interesting tweets [9], hashtags [15] and Twitter lists [43].

3 DATA COLLECTION

In this section, we describe our data collection strategy. Unlike the prior work [18] which collected only a small number of D-tweets (~4000), we employed the Hoaxy system [46] to collect

|D-tweets|	|S-tweets|	|D-guardians|	|S-guardians|	|D&S guardians|
157,482	67,586	70,900	45,406	7,167

Table 1: Statistics of our dataset.

Top15 D-guardians and # of D-tweets		
RandoRodeo (450)	stuartbirdman (318)	upayr (214)
pjr_cunningham (430)	ilpiese (297)	JohnOrJane (213)
TXDemocrat (384)	BreastsR4babies (255)	GreenPeaches2 (199)
Jkj193741 (355)	rankled2 (230)	spencerthayer (195)
BookRageStuff (325)	___lor__ (221)	SaintHeartwing (174)
Top 15 S-guaridans and # of S-tweets		
Jkj193741 (294)	MrDane1982 (49)	LeChatNoire4 (35)
MudNHoney (229)	pinch0salt (46)	bjcrochet (34)
_sirtainly (75)	ActualFlatticus (42)	upayr (33)
Paul197 (66)	BeltwayPanda (36)	58isthenew40 (33)
Endoracrat (49)	EJLandwehr (36)	slasher48 (31)

Table 2: Top 15 most active D-guardians and S-guardians, and associated # of D-tweets and # of S-tweets.

Verified guardians and (|D-tweets| vs. |S-tweets|)		
fawfulfan (103-1)	tomcoates (37-0)	KimLaCapria (27-3)
OpenSecretsDC (37-30)	**aravosis** (29-8)	PattyArquette (29-0)
PolitiFact (41-17)	**TalibKweli** (27-8)	NickFalacci (28-0)
RobertMaguire_ (46-7)	rolandscahill (31-0)	AaronJFentress (28-0)
jackschofield (42-1)	MichaelKors (30-0)	ParkerMolloy (26-1)

Table 3: Top 15 verified guardians, and corresponding D-tweet and S-tweet count.

a large number of both D-tweets and S-tweets. In particular, we collected 231,377 unique *fact-checking tweets* from six well-known fact-checking websites - *Snopes.com*, *Politifact.com*, *FactCheck.org*, *OpenSecrets.org*, *TruthOrfiction.com* and *Hoax-slayer.net* – via the APIs provided by the Hoaxy system which internally used Twitter streaming API. The collected data consisted of 161,981 D-tweets and 69,396 S-tweets (58,821 retweets of D-tweets and 10,575 quotes of D-tweets) generated from May 16, 2016 to July 7, 2017 (~ 1 year and 2 month). The number of our collected D-tweets is 40 times larger than the dataset used in the prior work [18].

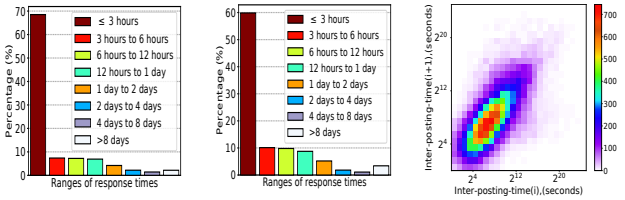
Similar to the prior work, we removed tweets containing only base URLs (e.g., snopes.com or politifact.com) or URLs simply pointing to the background information of the websites because the tweets containing these URLs may not reflect fact-checking enthusiasm and not contain fact-checking information. After filtering, we had 225,068 fact-checking tweets consisting of 157,482 D-tweets and 67,586 S-tweets posted by 70,900 D-guardians and 45,406 S-guardians. 7,167 users played both roles of D-guardians and S-guardians. The number of unique fact-checking URLs was 7,295. In addition, we also collected each guardian's recent 200 tweets. Table 1 shows the statistics of our pre-processed dataset.

4 CHARACTERISTICS OF GUARDIANS

From our dataset, we seek to answer the following research questions about guardians, their temporal behavior and topical interests.

Who are the guardians?

As we have shown in the previous section, there were only 7,167 users (7%) who behaved as both D-guardians and S-guardians, which indicates that guardians usually focused on either fact-checking claims in conversations (i.e., being D-guardians) or simply sharing



(a) D-guardians' response time (b) S-guardians' response time (c) S-tweets' inter-posting time

Figure 2: Ranges of response time of D-guardians and S-guardians, and inter-posting time of S-tweets. The color in (c) indicates the number of pairs.

credible information (i.e., being S-guardians). Since D-guardians and S-guardians played different roles, we seek to understand which group is more enthusiastic about its role. We created two lists - a list of the number of D-tweets posted by each D-guardian and a list of the number of S-tweets posted by each S-guardian –, excluding D&S guardians who performed both roles. Then, by conducting One-sided MannWhitney U-test, we found that D-guardians were significantly more enthusiastic about their role than S-guardians (p-value<10^{-6}). We also found that even the D&S guardians posted relatively larger number of D-tweets than S-tweets according to Wilcoxon one-sided test (p-value<10^{-6}).

The majority of guardians (85.3%) posted only 1~2 fact-checking tweets. However, there were super active guardians, each of whom posted over 200 fact-checking tweets. Table 2 shows the top 15 most active D-guardians and S-guardians and the number of their D-tweets and S-tweets. As we can see, the most active D-guardians showed their strong enthusiasm for posting fact-checked content in online discussions. Red-colored *Jkj193741* and *upayr* guardians were especially active in joining online conversations and spreading fact-checked information.

Next, we examined whether guardians have *verified* Twitter accounts or are highly visible users, who have at least 5,000 followers. The verified accounts and highly visible users usually play an important role in social media since their fact-checking tweets can reach many audiences [28, 47]. Since the verified accounts are more trustworthy, their fact-checking tweets are often shared by many other users. In our dataset, 2,401 guardians (2.2%) had verified accounts. Table 3 shows the top 15 verified accounts. Interestingly, some of these verified accounts behaved as D&S guardians, highlighted with the blue color in the table. Particularly, @PolitiFact, and @OpenSecretsDC, the official accounts of Politifact.com and OpenSecrets.org, frequently engaged in many online conversations. 8,221 guardians (7.5%) were highly visible users. Most top verified guardians, and many top S-guardians had a large number of followers. Altogether, S-tweets of the 45,406 S-guardians reached over 200 million followers.

Based on the analysis, we conclude that both D-guardians and S-guardians played important roles in terms of fact-checking claims and spreading the fact-checked news to the other users. Therefore, we need both types of guardians to spread credible information.

How quickly did guardians respond?

To further understand activeness of guardians, we examined how quickly D-guardians posted their fact-checking URLs as responses

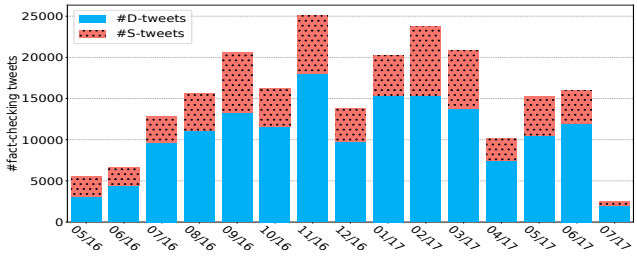


Figure 3: Temporal changes of #fact-checking tweets

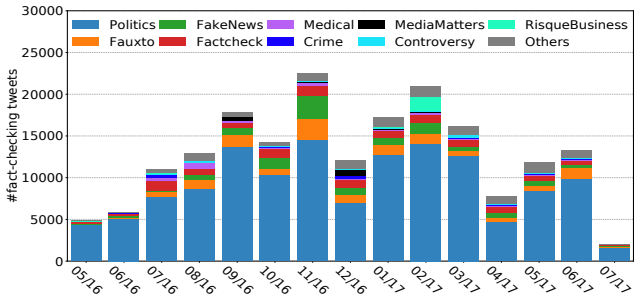


Figure 4: Topical changes of fact-checking tweets

to original posters' claims in online conversations. In particular, we measured response time of a D-tweet/D-guardian as a gap between an original poster's posting time and the fact-checking D-tweet's time. We collected all response time of D-tweets, grouped them and plotted a bar chart in Figure 2(a). The mean and median of response time were 2.26 days and 34 minutes, respectively. 90% of D-tweets were posted within one day, indicating D-guardians quickly responded to the claims and expressed their enthusiasm by posting fact-checking URLs/tweets.

Similarly, we also measured response time of an S-tweet/S-guardian (Figure 2(b)) as a gap between D-tweet's posting time and the corresponding S-tweet's posting time. The mean and median of the response time were 3.1 days and 90 minutes, respectively. 88.5% of S-tweets were posted within one day, indicating S-guardians also quickly responded and spread fact-checked information.

Finally, we measured S-guardians' inter-posting time to understand how long it took between two consecutive S-tweets, given the corresponding D-tweet. First, we grouped S-tweets based on each corresponding D-tweet, and sorted them in the ascending order of S-tweet creation time. Next, within each group, we computed inter-posting time δ_i as a gap between two consecutive S-tweets i and $i+1$ and created pairs of inter-posting time (δ_i, δ_{i+1}). These pairs were merged across all the groups and were plotted in log2 scale in Figure 2(c). Overall, the average inter-posting time was 5 minutes, which means an S-tweet was posted once per 5 minutes by S-guardians after the corresponding D-tweet was posted. To sum up, both D-guardians and S-guardians were active and quickly responded to claims and fact-checked content.

How did the volume of fact-checking tweets change over time? How did topics associated with fact-checking pages change over time?

First, we examined the change in the number of fact-checking tweets (i.e., D-tweets and S-tweets) in each month between May 2016 and July 2017. Figure 3 shows temporal changes of the number of fact-checking tweets. In the first 5 months, the number of fact-checking tweets increased gradually. In November 2016, the number of fact-checking tweets reached the peak (25,000 tweets) because of the US presidential election which happened on November 8, 2016. We also noticed that the number of D-tweets were larger than the number of S-tweets in every month which reflects that D-guardians were more active than S-guardians in online conversations (Wilcoxon one-side test p-value=3.052×10^{-5}). However, both D-guardians and S-guardians consistently posted and spread fact-checking tweets, respectively.

Next, we were interested in understanding what topics the fact-checking pages (linked by the URLs) were associated with and whether these topics changed over time. We first checked if a fact-checking website has categories, and if it did, we checked if we could automatically get the category information associated with each fact-checking page. For Snope pages, we identified each fact-checking page's topic by extracting the breadcrumb or tag information on the fact-checking page. We annotated PolitiFact pages' topic as *politics* due to its political missions. In this analysis, we did not include fact-checking pages associated with the other four fact-checking websites because there were no explicit categories in content of the fact-checking pages, and their coverage was only 17.22% (which would not contribute much to topical changes). Figure 4 shows temporal topical changes of fact-checking tweets in each month. Overall, *politics* was the most popular in all months. Interestingly, fact-checking tweets under *fauxtography*, *fake news* and *fact check* increased significantly in November 2016 (the month of US presidential election). In short, guardians' fact-checking activities were consistent over time, and their topical interests were mainly *politics*, *fauxtography* and *fake news*.

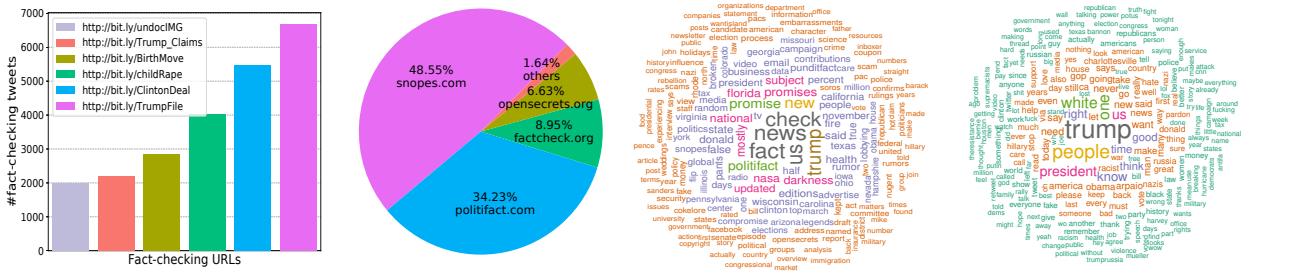
What fact-checking URLs were spread most by the guardians? What fact-checking websites did guardians embed in fact-checking tweets? What were the most important terms used in the fact-checking pages and 200 recent tweets?

Figure 5(a) shows the six most popular URLs embedded in fact-checking tweets. The URLs were related to Hillary Clinton and Donald Trump. Figure 5(b) shows what websites guardians used as references. Snopes.com was the most popular website, and politifact.com was the next frequently used one (48.55% vs. 34.23%).

To answer the last question, given a fact-checking page linked by each of D-tweets and S-tweets, we extracted the main content after removing headers, footers and irrelevant content. Then, we selected the top 250 words according to tf-idf values. Similarly, given 200 recent tweets of each guardian, we first aggregated them to make a big document, removed non-English tweets, stop words, and URLs. Then, we selected the top 250 words according to tf-idf values. As shown in Figure 5(c) and 5(d), "trump" was mentioned often in both word clouds. Surprisingly, "hillary" and "clinton" were less frequently mentioned than Trump-related words. The figures also confirm that politics were one of popular topics, especially Trump-related news was one of popular claims.

5 FACT-CHECKING URL RECOMMENDATION

In the previous section, we found that the guardians are enthusiastic about credibility of information on social network and highly



(a) Fact-checking URLs (b) Fact-checking websites (c) Top words in fact-checking pages (d) Top words in recent 200 tweets

Figure 5: (a) The most spread fact-checking URLs, (b) the most popular fact-checking websites, (c) the most important words in fact-checking pages linked by D-tweets and S-tweets and (d) the most important words in 200 recent tweets

active in spreading fact-checked content. To encourage them to further engage in disseminating verified information, we propose a recommendation model to personalize fact-checking URLs. The aim of the recommendation model is to help guardians quickly access new interesting fact-checking URLs/pages so that they could embed them in their messages, correct unverified claims or misinformation, and spread fact-checked information. We use terms "fact-checking URLs" and "URLs", interchangeably.

5.1 Problem Statement

Let $\mathcal{N} = \{u_1, u_2, ..., u_N\}$ and $\mathcal{M} = \{\ell_1, \ell_2, ..., \ell_M\}$ be a set of N guardians and a set of M fact-checking URLs, respectively. We view the action of embedding a fact-checking URL ℓ_j into a fact-checking tweet of guardian u_i as an interaction pair (u_i, ℓ_j). We form a matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$ where $\mathbf{X}_{ij} = 1$ if the guardian u_i posted a fact-checking URL ℓ_j. Otherwise, $\mathbf{X}_{ij} = 0$. Our main goal is to learn a model that recommends similar URLs to guardians whose interests are similar. In particular, we aim to learn matrix $\mathbf{U} \in \mathbb{R}^{N \times D}$, where each row vector $U_i^T \in \mathbb{R}^{D \times 1}$ is the latent representation of guardian u_i, and matrix $\mathbf{V} \in \mathbb{R}^{D \times M}$, where each column vector $V_j \in \mathbb{R}^{D \times 1}$ is the latent representation of URL ℓ_j. $D \ll min(M, N)$ is latent dimensions. Toward the goal, we propose our initial/basic matrix factorization model as follows:

$$\min_{\mathbf{U},\mathbf{V}} \|\Omega \odot (\mathbf{X} - \mathbf{UV})\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) \tag{1}$$

where $\Omega \in \mathbb{R}^{N \times M}$, and $\Omega_{ij} = 1$ if $\mathbf{X}_{ij} = 1$. Otherwise, $\Omega_{ij} = 0$. Operators $\odot$ and $\|.\|_F^2$ are Hadamard product and Frobenius norm, respectively. Finally, λ is regularization factor to avoid overfitting.

5.2 Co-ocurrence model

Now, we turn to extend our basic model in Eq.1 by further utilizing the interaction matrix $\mathbf{X}$. Inspired by [32, 39], we propose to regularize our basic model in Eq.1 by generating two additional matrices - URL-URL co-occurrence matrix and guardian-guardian co-occurrence matrix. Our main intuition of the extension is that a pair of URLs, which were posted by the same guardian, may be similar to each other. Likewise, a pair of guardians who posted the same URLs may be alike. To better understand our proposed models, we present the word embedding model as background information.

5.2.1 Word embedding model. Given a sequence of training words, word embedding models attempt to learn the distributed vector representation of each word. A typical example is *word2vec* proposed by Mikolov et al. [39]. Given a training word w, the main objective of the skip-gram model in *word2vec* is to predict the *context words* (i.e. the words that appear in a fixed-size context window) of w. Recently, it has been shown that training skip-gram model with negative sampling is similar to factorizing a word-context matrix named Shifted Positive Pointwise Mutual Information matrix (*SPPMI*) [30]. Given a word i and its context word j, the value $SPPMI(i, j)$ is computed as follows:

$$SPPMI(i, j) = max\{PMI(i, j) - log(s), 0\} \tag{2}$$

where $s \geq 1$ is the number of negative samples, and $PMI(i, j)$ is an element of Pointwise Mutual Information (PMI) matrix. $PMI(i, j)$ is estimated as $\log\left(\frac{\#(i,j) \times |D|}{\#(i) \times \#(j)}\right)$ where $\#(i, j)$ is the number of times that word j appears in the context window of word i. $\#(i) = \sum_j \#(i, j)$, and $\#(j) = \sum_i \#(i, j)$. $|D|$ is the total number of pairs of word and context word. Note that $PMI(i, i) = 0$ for every word i.

5.2.2 URL-URL co-occurrence. We generate a matrix $\mathbf{R} \in \mathbb{R}^{M \times M}$ where $\mathbf{R}_{ij} = SPPMI(\ell_i, \ell_j)$ based on co-occurrence of URLs. In particular, for each URL ℓ_i posted by a specific guardian, we define its context as all other URLs ℓ_j posted by the same guardian. Based on this definition, $\#(i, j)$ means the number of guardians that posted both URL ℓ_i and ℓ_j. $\#(i, j)$ is also interpreted as the co-occurrence of URL ℓ_i and URL ℓ_j. After that, we compute $PMI(\ell_i, \ell_j)$ and $SPPMI(\ell_i, \ell_j)$ based on Equation 2 for all pairs of ℓ_i and ℓ_j.

5.2.3 Guardian-Guardian co-occurrence. Similarly, the context for each guardian u_i is defined as all other guardians u_j who posted the same URL with u_i. Then, $\#(i, j)$ is the number of URLs that both guardian u_i and guardian u_j commonly posted. Given this definition, we can generate a SPPMI matrix $\mathbf{G} \in \mathbb{R}^{N \times N}$ where $\mathbf{G}_{ij} = SPPMI(u_i, u_j)$. The same value of hyper-parameter s is used for generating matrices $\mathbf{R}$ and $\mathbf{G}$.

5.2.4 Regularizing matrix factorization with co-occurrence matrices. Our intuition is that URLs which are commonly posted by similar set of guardians are similar, and guardians who commonly posted the same set of URLs are close to each other. With that intuition, we propose loss function $\mathcal{L}_{XRG}$ – a joint matrix factorization model of three matrices $\mathbf{X}$, $\mathbf{R}$ and $\mathbf{G}$ as follows:

$$\begin{aligned}\mathcal{L}_{XRG} = &\|\Omega \odot (\mathbf{X} - \mathbf{UV})\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) \\ &+ \|\mathbf{R}^{mask} \odot (\mathbf{R} - \mathbf{V}^T\mathbf{K})\|_F^2 + \|\mathbf{G}^{mask} \odot (\mathbf{G} - \mathbf{UL})\|_F^2\end{aligned} \tag{3}$$

where $\mathbf{R}^{mask} \in \mathbb{R}^{M \times M}$, $\mathbf{R}^{mask}_{ij} = 1$ if $\mathbf{R}_{ij} > 0$. Otherwise, $\mathbf{R}^{mask}_{ij} = 0$. $\mathbf{G}^{mask} \in \mathbb{R}^{N \times N}$, $\mathbf{G}^{mask}_{ij} = 1$ if $\mathbf{G}_{ij} > 0$. Otherwise, $\mathbf{G}^{mask}_{ij} = 0$. Two matrices $\mathbf{K} \in \mathbb{R}^{D \times M}$ and $\mathbf{L} \in \mathbb{R}^{D \times N}$ act as additional parameters. Although our work shares similar ideas with [32], there are three key differences between our model and [32] as follows: (1) we omit bias matrices to reduce model complexity which is helpful in reducing overfitting, (2) additional matrix $\mathbf{G}$ is factorized and (3) we do not regularize parameters $\mathbf{K}$ and $\mathbf{L}$.

5.3 Integrating Auxiliary Information

In addition, we propose auxiliary information which will be integrated with Eq.3 to improve URL recommendation performance.

5.3.1 Modeling social structure. The social structure of guardians may reflect the homophily phenomenon indicating that guardians who follow each other may have similar interests in fact-checking URLs [50]. To model this social structure of guardians, we first construct an unweighted undirected graph $G(V, E)$ where nodes are guardians, and an edge (u_i, u_j) between guardians u_i and u_j are formed if u_i follows u_j or u_j follows u_i. In our dataset, in total, there were 1,033,704 edges in $G(V, E)$ (density=0.013898), which is 5.9 times higher than reported density in [57], indicating dense connections between guardians. We represent $G(V, E)$ by using an adjacency matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ where $\mathbf{S}_{ij} = 1$ if there is an edge (u_i, u_j). Otherwise, $\mathbf{S}_{ij} = 0$. Second, we use Equation 4 as a regularization term to make latent representations of connected guardians similar to each other. Then, we formally minimize $\mathcal{L}_1$ as follows:

$$\mathcal{L}_1 = \|\mathbf{S} - \mathbf{U}\mathbf{U}^T\|_F^2 \tag{4}$$

5.3.2 Modeling topical interests based on 200 recent tweets. In addition to social structure, the content of 200 recent tweets may reflect guardians' interests [1, 2, 8]. In Figure 5(d), 200 recent tweets of guardians contain many political words, which suggests us to enrich guardians' latent representation based on tweets' content.

For each guardian, we build a document by aggregating his/her 200 recent tweets and then employ the Doc2Vec model [26] to learn latent representations of the document. Doc2Vec is an unsupervised learning algorithm, which automatically learns high quality representation of documents. We use Gensim[4] as implementation of the Doc2Vec, set 300 as latent dimensions of documents, and train Doc2Vec model for 100 iterations. After training Doc2Vec model, we derive cosine similarity of every pair of learned vectors to create a symmetric matrix $\mathbf{X}_{uu} \in \mathbb{R}^{N \times N}$, where $\mathbf{X}_{uu}(i, j) \in [0; 1]$ represents the similarity of document vectors of guardians u_i and u_j. Intuitively, if two guardians have similar interests, their document vectors may be similar. Thus, we regularize guardians' latent representations to make them as close as possible by minimizing the following objective function:

$$\begin{aligned}\mathcal{L}_2 &= \frac{1}{2}\sum_{i=1,j=1}^{N} \mathbf{X}_{uu}(i,j)\|U_i^T - U_j^T\|^2 \\ &= \sum_{i=1}^{N} U_i^T \mathbf{D}_{uu}(i,i) U_i - \sum_{i=1,j=1}^{N} U_i^T \mathbf{X}_{uu}(i,j) U_j \\ &= Tr(\mathbf{U}^T \mathbf{D}_{uu}\mathbf{U}) - Tr(\mathbf{U}^T \mathbf{X}_{uu}\mathbf{U}) = Tr(\mathbf{U}^T \mathcal{L}_{uu}\mathbf{U})\end{aligned} \tag{5}$$

[4] https://radimrehurek.com/gensim/

where $\mathbf{D}_{uu} \in \mathbb{R}^{N \times N}$ is a diagonal matrix with diagonal element $\mathbf{D}_{uu}(i, i) = \sum_{j=1}^{N} \mathbf{X}_{uu}(i, j)$. $Tr(.)$ is the trace of matrix, and $\mathcal{L}_{uu} = \mathbf{D}_{uu} - \mathbf{X}_{uu}$, which is a Laplacian matrix of the matrix $\mathbf{X}_{uu}$.

5.3.3 Modeling topical similarity of fact-checking pages. We further exploit the content of fact-checking URLs (i.e., fact-checking pages) as an additional data source to improve recommendation quality. As we can see in Figure 5(c), URLs' contents are mostly about politics. Intuitively, if the content of two URLs are similar (e.g. they are about Hillary Clinton's foundation as shown in Figure 1), their latent representations should be close. Exploiting the content of a fact-checking URL has been employed in [2, 51]. In this paper, we apply a different approach, in which the Doc2Vec model is utilized to learn latent representation of URLs. Hyperparameters of the Doc2Vec model are the same as what we used for content of tweets. After training the Doc2Vec model, we derive the symmetric similarity matrix $\mathbf{X}_{\ell\ell} \in \mathbb{R}^{M \times M}$ and minimize the loss function $\mathcal{L}_3$ in Equation 6 as a way to regulate latent representation of URLs.

$$\begin{aligned}\mathcal{L}_3 &= \frac{1}{2}\sum_{i=1,j=1}^{M} \mathbf{X}_{\ell\ell}(i,j)\|V_i - V_j\|^2 \\ &= \sum_{i=1}^{M} V_i \mathbf{D}_{\ell\ell}(i,i) V_i^T - \sum_{i=1,j=1}^{M} V_i \mathbf{X}_{\ell\ell}(i,j) V_j^T \\ &= Tr(\mathbf{V}(\mathbf{D}_{\ell\ell} - \mathbf{X}_{\ell\ell})\mathbf{V}^T) \\ &= Tr(\mathbf{V}\mathcal{L}_{\ell\ell}\mathbf{V}^T)\end{aligned} \tag{6}$$

where $\mathbf{D}_{\ell\ell} \in \mathbb{R}^{M \times M}$ is a diagonal matrix with elements on the diagonal $\mathbf{D}_{\ell\ell}(i, i) = \sum_{j=1}^{M} \mathbf{X}_{\ell\ell}(i, j)$ and $\mathcal{L}_{\ell\ell} = \mathbf{D}_{\ell\ell} - \mathbf{X}_{\ell\ell}$, which is the graph Laplacian of the matrix $\mathbf{X}_{\ell\ell}$.

5.4 Joint-learning fact-checking URL recommendation model

Finally, we propose GAU - a joint model of **G**uardian-Guardian SPPMI matrix, **A**uxiliary information and **U**RL-URL SPPMI matrix. The objective function of our model, $\mathcal{L}_{GAU}$, is presented in Eq.7:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V},\mathbf{L},\mathbf{K}} \mathcal{L}_{GAU} &= \|\Omega \odot (\mathbf{X} - \mathbf{U}\mathbf{V})\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) \\ &+ \|\mathbf{R}^{mask} \odot (\mathbf{R} - \mathbf{V}^T\mathbf{K})\|_F^2 \\ &+ \|\mathbf{G}^{mask} \odot (\mathbf{G} - \mathbf{U}\mathbf{L})\|_F^2 \\ &+ \alpha \times \|\mathbf{S} - \mathbf{U}\mathbf{U}^T\|_F^2 \\ &+ \gamma \times Tr(\mathbf{U}^T\mathcal{L}_{uu}\mathbf{U}) \\ &+ \beta \times Tr(\mathbf{V}\mathcal{L}_{\ell\ell}\mathbf{V}^T)\end{aligned} \tag{7}$$

where $\alpha, \gamma, \beta, \lambda$ and shifted negative sampling value s are hyper parameters, tuned based on a validation set. We optimize $\mathcal{L}_{GAU}$ by using gradient descent to iteratively update parameters with fixed learning rate $\eta = 0.001$. The details of the optimization algorithm are presented in Algorithm 1. After learning $\mathbf{U}$ and $\mathbf{V}$, we estimate the guardian u_i's preference for URL ℓ_j as: $\hat{r}_{i,j} \approx U_i V_j$. The final URLs recommended for a guardian u_i is formed based on ranking:

$$u_i : \ell_{j_1} > \ell_{j_2} > \ldots > \ell_{j_M} \rightarrow \hat{r}_{i,j_1} > \hat{r}_{i,j_2} > \ldots > \hat{r}_{i,j_M} \tag{8}$$

The derivatives of loss $\mathcal{L}_{GAU}$ with respect to parameters $\mathbf{U}$, $\mathbf{V}$, $\mathbf{K}$ and $\mathbf{L}$ are as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{U}} &= -2(\Omega \odot \Omega \odot (\mathbf{X} - \mathbf{UV}))\mathbf{V}^T + 2\lambda \times (\mathbf{U}) \\
&\quad -2(\mathbf{G}^{mask} \odot \mathbf{G}^{mask} \odot (\mathbf{G} - \mathbf{UL}))\mathbf{L}^T \\
&\quad -2\alpha((\mathbf{S} - \mathbf{UU}^T + (\mathbf{S} - \mathbf{UU}^T)^T)\mathbf{U}) \\
&\quad +\gamma \times (\mathcal{L}_{uu} + \mathcal{L}_{uu}^T)\mathbf{U} \\
\frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{V}} &= -2\mathbf{U}^T(\Omega \odot \Omega \odot (\mathbf{X} - \mathbf{UV})) + 2\lambda \times (\mathbf{V}) \\
&\quad -2\mathbf{K}(\mathbf{R}^{mask} \odot \mathbf{R}^{mask} \odot (\mathbf{R} - \mathbf{V}^T\mathbf{K}))^T \\
&\quad +\beta \times \mathbf{V}(\mathcal{L}_{\ell\ell} + \mathcal{L}_{\ell\ell}^T) \\
\frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{L}} &= -2\mathbf{U}^T(\mathbf{G}^{mask} \odot \mathbf{G}^{mask} \odot (\mathbf{G} - \mathbf{UL})) \\
\frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{K}} &= -2\mathbf{V}(\mathbf{R}^{mask} \odot \mathbf{R}^{mask} \odot (\mathbf{R} - \mathbf{V}^T\mathbf{K}))
\end{aligned} \tag{9}$$

Algorithm 1 GAU Optimization algorithm

Input: Guardian-URL interaction matrix $\mathbf{X}$, URL-URL SPPMI matrix $\mathbf{R}$, Guardian-Guardian SPPMI matrix $\mathbf{G}$, social structure matrix $\mathbf{S}$, Laplacian matrix $\mathcal{L}_{uu}$ of guardians, Laplician matrix $\mathcal{L}_{\ell\ell}$ of URLs, binary matrices Ω, $\mathbf{R}^{mask}$ and $\mathbf{G}^{mask}$ as indication matrices.
Output: $\mathbf{U}$ and $\mathbf{V}$

Initialize $\mathbf{U}$, $\mathbf{V}$, $\mathbf{K}$ and $\mathbf{L}$ with Gaussian distribution $\mathcal{N}(0, 0.01^2)$, $t \leftarrow 0$
while Not Converged **do**
  Compute $\frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{U}}$, $\frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{V}}$, $\frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{L}}$ and $\frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{K}}$ in Eq.9
  $\mathbf{U}_{t+1} \leftarrow \mathbf{U}_t - \eta \frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{U}}$
  $\mathbf{V}_{t+1} \leftarrow \mathbf{V}_t - \eta \frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{V}}$
  $\mathbf{L}_{t+1} \leftarrow \mathbf{L}_t - \eta \frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{L}}$
  $\mathbf{K}_{t+1} \leftarrow \mathbf{K}_t - \eta \frac{\partial \mathcal{L}_{GAU}}{\partial \mathbf{K}}$
  $t \leftarrow t + 1$
return $\mathbf{U}$ and $\mathbf{V}$

6 EVALUATION

In this section, we thoroughly experiment our proposed *GAU* model. In particular, we aim to answer the following research questions:

- **RQ1:** What is the benefit of integrating auxiliary data such as tweets, fact-checking URL's content and network structure?
- **RQ2:** How helpful is adding SPPMI matrices of fact-checking URLs and guardians?
- **RQ3:** What is the performance of the proposed *GAU* model compared with other state-of-the-arts methods?
- **RQ4:** What is the performance of the proposed GAU model for different types of guardians in terms of activeness level?
- **RQ5:** What is the sensitivity of *GAU* to hyperparameters?

6.1 Experimental Settings

Processing our dataset. We were interested in selecting active and professional guardians who frequently posted fact-checking URLs since they would be more likely to spread recommended fact-checking URLs than casual guardians.

Following a similar preprocessing approach to recommending scientific articles [51, 52], we only selected guardians who used at least three distinct fact-checking URLs in their D-tweets and/or S-tweets. Altogether, 12,197 guardians were selected for training and evaluating recommendation models. They posted 4,834 distinct fact-checking URLs in total. The number of interactions was 68,684 (Sparsity:99.9%). There were 9,710 D-guardians, 6,674 S-guardians and 4,187 users who played both roles. The total number of followers of the 12,197 guardians was 55,325,364, indicating their high impact on fact-checked information propagation.

Experimental design and metrics. To validate our model, we followed a similar approach that [32] did. In particular, we randomly selected 70%, 10% and 20% URLs of each guardian for training, validation and testing. The validation data was used to tune hyperparameters and to avoid overfitting. We repeated this evaluation scheme for five times, getting five different sets of training, validation and test data. The average results were reported. We used three standard ranking metrics such as Recall@k, MAP@k (Mean Average Precision) and NDCG@k (Normalized Discounted Cumulative Gain) [32, 37]. Since $k = 10$ was used in [1], we tested our model with $k \in \{5, 10, 15\}$.

6.2 Baselines and Our Model

We compared our proposed model with the following four state-of-the-art collaborative filtering algorithms:

- **BPRMF** Bayesian Personalized Ranking Matrix Factorization [44] optimizes the matrix factorization model with pairwise ranking loss. It is a common baseline for item recommendation.
- **MF** Matrix Factorization (MF) [22] is a standard technique in collaborative filtering. Given an interaction matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$, it factorizes $\mathbf{X}$ into two matrices $\mathbf{U} \in \mathbb{R}^{N \times D}$ and $\mathbf{V} \in \mathbb{R}^{D \times M}$, which are latent representations of users and items, respectively.
- **CoFactor** CoFactor [32] extended Weighted Matrix Factorization (WMF) by jointly decomposing interaction matrix $\mathbf{X}$ and co-occurrence SPPMI matrix for items (i.e., fact-checking URLs in this context). We set a confidence value $c_{X_{ij}=1} = 1.0$ for $X_{ij} = 1$, and we set $c_{X_{ij}=0} = 0.01$ for non-observed interaction. The number of negative samples s was grid-searched in a set $s \in \{1, 2, 5, 10, 50\}$, following the same settings as in [32].
- **CTR** Collaborative Filtering Regression [51] employed content of URLs (i.e., fact-checking pages in this context) to recommend scientific papers to users. Following exactly the best setting reported in the paper, we selected the top 8,000 words from fact-checking URLs' contents based on the mean of tf-idf values and set $\lambda_u = 0.01$, $\lambda_v = 100$, D=200, a=1 and b=0.01.

To build our GAU model, we conducted the grid-search to select the best value of α, β and γ in $\{0.02, 0.04, 0.06, 0.08\}$. The number of negative samples s for constructing SPPMI matrices was in $\{1, 2, 5, 10, 50\}$. For all of the baselines and the GAU model, we set latent dimensions to $D = 100$ unless explicitly stated, and regularization value λ was grid-searched in $\{10^{-5}, 3\times10^{-5}, 5\times10^{-5}, 7\times10^{-5}\}$ by default. We only report the best result of each baseline.

We also attempted to compare our proposed model with content-based recommendation algorithms [1–3, 56]. These methods mostly required collecting additional data from external data sources which are very time-consuming and expensive, and sometimes impossible for the third party researchers. We tried to compare our model with

Methods	Recall@5	NDCG@5	MAP@5	Recall@10	NDCG@10	MAP@10	Recall@15	NDCG@15	MAP@15	Avg. Rank
BASIC	0.08919 (6)	0.06004 (6)	0.04839 (6)	0.13221 (6)	0.07417 (6)	0.05413 (6)	0.16208 (6)	0.08227 (6)	0.05653 (6)	6.0
BASIC+NW+UC	0.09967 (4)	0.06814 (5)	0.05535 (5)	0.14817 (4)	0.08399 (4)	0.06170 (5)	0.18280 (3)	0.09335 (5)	0.06432 (5)	4.4
BASIC+NW+UC+CSU	0.09900 (5)	0.06822 (4)	0.05604 (4)	0.14688 (5)	0.08386 (5)	0.06235 (4)	0.18266 (4)	0.09354 (4)	0.06522 (4)	4.3
BASIC+CSU+CSG	0.10247 (3)	0.06958 (3)	0.05670 (3)	0.14950 (3)	0.08497 (3)	0.06293 (3)	0.18205 (5)	0.09380 (3)	0.06554 (3)	3.2
BASIC+NW+UC+CSU+CSG	0.11133 (2)	0.07422 (2)	0.05978 (2)	0.16127 (2)	0.09065 (2)	0.06646 (2)	0.19516 (2)	0.09980 (2)	0.06917 (2)	2.0
Our GAU model	0.11582 (1)	0.07913 (1)	0.06481 (1)	0.16400 (1)	0.09489 (1)	0.07118 (1)	0.19693 (1)	0.10381 (1)	0.07382 (1)	1.0

Table 4: Effectiveness of using auxiliary information and co-occurrence matrices. The GAU model outperforms the other variants significantly with p-value<0.001.

recent work [56] and collected 5,383,598 followees of the 12,196 guardians and over 15 million distinct Twitter lists in which at least one of the followees was included. However, we were not able to collect all fact-checking tweets posted by these followees during the same data collection period (from May 16, 2016 to July 7, 2017). Therefore, we only used followees that were in the set of 12,197 guardians. But, maybe because of the limited data, it performed poorly in the experiments. Therefore, we omit its results in the experiments. Instead, we report performance of our GAU model and the four state-of-the-art collaborative filtering algorithms.

6.3 Effectiveness of Auxiliary Information and SPPMI Matrices (RQ1 & RQ2)

Before comparing our GAU model with the four baselines, we first examined the effectiveness of exploiting auxiliary information and the utility of jointly factorizing SPPMI matrices. Starting from our basic model in Eq.1, we created variants of the *GAU* model. Since there are many variants of *GAU*, we selectively report performance of the following *GAU*'s variants:

- Our basic model (Equation 1) (BASIC)
- BASIC + Network + URL's content (BASIC+NW+UC)
- BASIC + Network + URL's content + URL's SPPMI matrix (BASIC+NW+UC+CSU)
- BASIC + URL's SPPMI matrix + Guardians' SPPMI matrix (BASIC+CSU+CSG)
- BASIC + Network + URL's content + SPPMI matrix of URLs + SPPMI matrix of Guardians (BASIC+NW+UC+CSU+CSG)
- Our GAU model

Table 4 shows performance of the variants and the GAU model. It shows the rank of each method based on reported metrics. By adding social network information and fact-checking URL's content to Equation 1, there was a huge climb in performance of BASIC+NW+UC over BASIC across all metrics. In particular, Recall, NDCG and MAP of BASIC+NW+UC were better than BASIC about 12.20%±1.31%, 13.39%±0.34% and 14.04%±0.76%, respectively (confidence interval 95%). These results confirm the effectiveness of exploiting auxiliary information.

How about using co-occurrence SPPMI matrices of fact-checking URLs and guardians? First, when adding co-occurrence SPPMI matrix of fact-checking URL (CSU) to the variant BASIC+NW+UC, we did not see much improvement across all settings. Second, when jointly factorizing two SPPMI matrices (BASIC+CSU+CSG) and comparing it with the variant BASIC+NW+UC, we can see that BASIC+CSU+CSG and BASIC+NW+UC performed equally well. Again, BASIC+CSU+CSG did not use any additional data sources except the interaction matrix $\mathbf{X}$. It is an attractive benefit since it did not depend on other data sources. In other words, it reflects that regularizing the BASIC model with SPPMI matrices is comparable to adding network data and URLs' contents to the BASIC model.

So far, both auxiliary information and SPPMI matrices are beneficial to improving recommendation quality. How about combining all of them into a single model? Will performance be further improved? We turned to the variant BASIC+NW+UC+CSU+CSG. As expected, BASIC+NW+UC+CSU+CSG enhanced CSU+CSG by 7.90%±1.79% Recall, 6.58%±0.40% NDCG, and 5.53%±0.22% MAP. Its results were also higher than BASIC+NW+UC about 9.10%±6.15% Recall, 7.92%±2.50% NDCG and 7.75%±0.58% MAP.

Since adding auxiliary data was valuable, we now exploit another data source – 200 recent tweets' content. Consistently, adding the tweets' content indeed improved performance. The improvement of the GAU over BASIC+NW+UC+CSU+CSG model was 4.0% Recall, 6.6% NDCG and 8.4% MAP. This improvement is statistically significant with p-value<0.001 using Wilcoxon one-sided test. Comparing the GAU with the BASIC model, we observed a dramatic increase in performance across all metrics. Specifically, Recall, NDCG and MAP were improved by 25.13%±10.64%, 28.64%±7.13% and 32%±4.29% respectively.

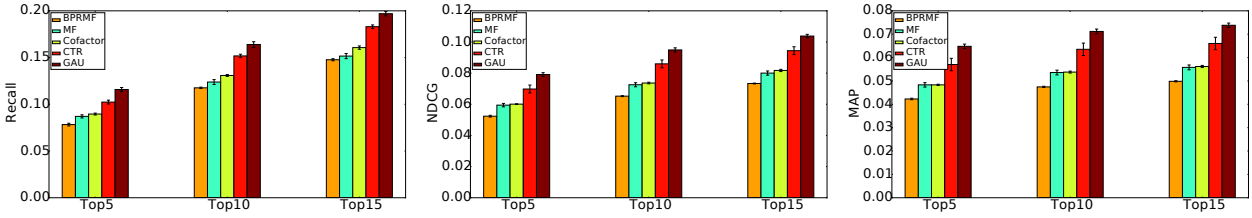
Based on the experiments, we conclude that auxiliary data as well as co-occurrence matrices are helpful to improve recommendation quality. Adding CSU+CSG or NW+UC enhanced the BASIC model by 12% to 14%. Our GAU model performed best, which improved the BASIC model by 25%∼32%.

6.4 Performance of GAU and Baselines (RQ3)

Figure 6 shows the performance of the four baselines and GAU. MF was better than BPRMF which was designed to optimize Area Under Curve (AUC). Similar results were reported in [55]. CTR was a very competitive baseline. This reflects the importance of fact-checking URL's content (i.e., fact-checking page) in recommending right fact-checking URLs to guardians. GAU performed better than CTR by 12.75% ± 0.95% Recall, 11.2% ± 4.6% NDCG, and 12.5% ± 2.5% MAP. GAU also outperformed CoFactor with a large margin by 25.8% ± 8.4% Recall, 29.2% ± 5.8% NDCG, and 32.6% ± 3.4% MAP (confidence interval 95%). Overall, our GAU model significantly outperformed all the baselines (p-value<0.001). The improvement over the baselines was 11%∼33%.

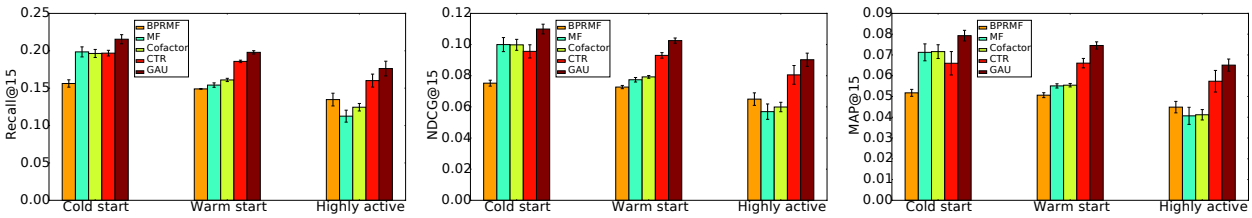
6.5 Performance of Models for Different Types of Guardians (RQ4)

We grouped guardians into three types based on the number of their fact-checking URLs (i.e., the activeness level) to see whether our GAU still outperforms the baselines in all the three types. By sorting guardians in the ascending order of the number of their



(a) Recall@k (b) NDCG@k (c) MAP@k

Figure 6: Performance of our GAU model and 4 baselines. The GAU model outperforms the baselines (p-value<0.001).



(a) Recall@15 (b) NDCG@15 (c) MAP@15

Figure 7: Performance of GAU and baselines for three types of guardians. GAU outperforms the baselines (p-value<0.01).

fact-checking URLs, we annotated the first 20% guardians as cold-start guardians, the next 60% guardians as warm-start guardians, and the last 20% guardians as highly active guardians.

Figure 7 shows performance of GAU and the baselines in Top 15 results. A general pattern of all the models was that they performed pretty well for cold-start guardians, and their performance slightly decreased as guardians posted more fact-checking URLs. We observed consistent results in top 5 and top10 as well.

GAU outperformed CTR in cold-start, warm-start and highly active guardians, improving Recall@15 by 6.5%∼10.0%, NDCG@15 by 10.2%∼15.0%, and MAP@15 by 12.8%∼20.1%. Overall, GAU consistently outperformed the baselines for all three groups according to the three metrics. Its improvement was about 6.5%∼20.1%.

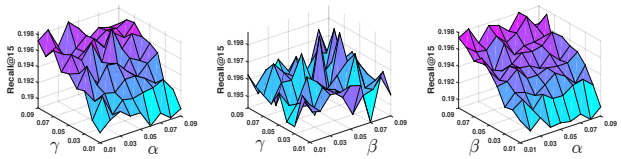
6.6 Exploiting hyper-parameters (RQ5)

We investigated the impact of hyper-parameters α, β and γ on the GAU model. These hyper-parameters control the contribution of social network, fact-checking URL's content and 200 recent tweets' content to the GAU. We tested α, β and γ from 0.01 to 0.09, increasing 0.01 in each step, and then report the average recall@15, while we fixed $\lambda = 3 \times 10^{-5}$ and the number of negative samples $s = 10$.

In Figure 8(a), we fixed $\beta = 0.08$ and varied α and γ. The general trend was that recall@15 gradually went up, when α and γ increased. It reached the peak, when $\alpha = 0.06$ and $\gamma = 0.06$. Next, we fixed $\alpha = 0.08$. It seems recall@15 fluctuated when varying β and γ, but the amplitude was small. The max Recall@15 was only 2.2% larger than the smallest Recall@15. Finally, γ was fixed to 0.08. The trend was similar to Figure 8(a). In general, when α, β and γ are large, the performance tends to improve, which suggests the importance of regularizing our model using the auxiliary information.

7 DISCUSSION

In Section 4, we showed that guardians had great enthusiasm for information credibility. Nevertheless, many guardians only posted 1∼2 fact-checking tweets. Therefore, we only recommended URLs



(a) $\beta = 0.08$ (b) $\alpha = 0.08$ (c) $\gamma = 0.08$

Figure 8: Hyper-parameter sensitivity.

to highly enthusiastic guardians, who posted at least 3 fact-checking URLs, because they may continue to be active in spreading fact-checked information in the future. Another observation is that the top verified guardians seem not to be active in the covered time period. We conjecture that these verified guardians may be cautious about what they should post to their followers [38]. We also showed that exploiting auxiliary information indeed helped improve recommendation quality. There is considerable potential to integrate other data sources such as temporal factors and activeness of guardians to further improve the proposed recommender system. We leave them for future work.

8 CONCLUSION

We collected a list of guardians, who showed their interests in information credibility by embedding fact-checking URLs in their posts. The guardians were very active in posting credible information and were mostly interested in politics, fauxotography and fake news. After analyzing our dataset, we proposed a recommendation model to personalize fact-checking URLs to the guardians toward enhancing their engagement in fact-checking activities and encouraging them to post more credible information. Our proposed model outperformed four baselines (i.e., MF, CoFactor, BPRMF and CTR). In future work, we will upgrade our model to address the cold-start issue where guardians posted less than 3 fact-checking URLs and will investigate whether employing deep learning techniques would further improve performance of our model.

ACKNOWLEDGMENT

This work was supported in part by NSF grant CNS-1755536, Google Faculty Research Award, and Microsoft Azure Research Award. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2011. Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *WebSci*.
[2] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2011. Semantic enrichment of twitter posts for user profile construction on the social web. In *Extended Semantic Web Conference*.
[3] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2013. Twitter-Based User Modeling for News Recommendations.. In *IJCAI*.
[4] Hunt Allcott and Matthew Gentzkow. 2017. *Social media and fake news in the 2016 election.* Technical Report. National Bureau of Economic Research.
[5] Samantha Bradshaw and Philip N Howard. 2017. Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. *University of Oxford, Computional Propaganda Research Project* (2017).
[6] Michael J Brzozowski and Daniel M Romero. 2011. Who Should I Follow? Recommending People in Directed Social Networks.. In *ICWSM*.
[7] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *WWW*.
[8] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. 2010. Short and tweet: experiments on recommending content from information streams. In *CHI*.
[9] Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. 2012. Collaborative personalized tweet recommendation. In *SIGIR*.
[10] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. 2012. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *WSDM*.
[11] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. 2010. Time is of the essence: improving recency ranking using twitter data. In *WWW*.
[12] Ullrich KH Ecker, Stephan Lewandowsky, and David TW Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition* 38, 8 (2010), 1087–1100.
[13] William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *NAACL HLT*.
[14] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor Cascades.. In *ICWSM*.
[15] Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. 2013. Using topic models for twitter hashtag recommendation. In *WWW*.
[16] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *WWW*.
[17] Ido Guy, Inbal Ronen, and Ariel Raviv. 2011. Personalized activity streams: sifting through the river of news. In *RecSys*.
[18] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations.. In *ICWSM*.
[19] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *KDD*.
[20] Muneo Kaigo. 2012. Social media usage during disasters and social capital: Twitter and the Great East Japan earthquake. *Keio Communication Review* 34, 1 (2012).
[21] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *WSDM*. ACM.
[22] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009).
[23] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Aspects of rumor spreading on a microblog network. In *SocInfo*.
[24] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *ICDM*.
[25] Zignal Labs. 2017. A Report on the Spread of Fake News. http://go.zignallabs.com/Q1-2017-fake-news-report. (2017).
[26] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*.
[27] Kyumin Lee, Brian David Eoff, and James Caverlee. 2011. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter.. In *ICWSM*.
[28] Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle Zhou, and Jeffrey Nichols. 2014. Who will retweet this?: Automatically identifying and engaging strangers on twitter to spread information. In *IUI*.
[29] Kyumin Lee, Prithivi Tamilarasan, and James Caverlee. 2013. Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media.. In *ICWSM*.
[30] Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *NIPS*.
[31] Quanzhi Li, Xiaomo Liu, Rui Fang, Armineh Nourbakhsh, and Sameena Shah. 2016. User Behaviors in Newsworthy Rumors: A Case Study of Twitter.. In *ICWSM*.
[32] Dawen Liang, Jaan Altosaar, Laurent Charlin, and David M Blei. 2016. Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence. In *RecSys*.
[33] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *CIKM*.
[34] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks.. In *IJCAI*.
[35] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *CIKM*.
[36] Jim Maddock, Kate Starbird, Haneen J Al-Hassani, Daniel E Sandoval, Mania Orand, and Robert M Mason. 2015. Characterizing online rumoring behavior using multi-dimensional signatures. In *CSCW*.
[37] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge University Press.
[38] Alice Marwick and Danah Boyd. 2011. To see and be seen: Celebrity practice on Twitter. *Convergence* 17, 2 (2011), 139–158.
[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.
[40] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.
[41] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *WWW*.
[42] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *EMNLP*.
[43] Vineeth Rakesh, Dilpreet Singh, Bhanukiran Vinzamuri, and Chandan K Reddy. 2014. Personalized Recommendation of Twitter Lists using Content and Network Information.. In *ICWSM*.
[44] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Uncertainty in Artificial Intelligence*.
[45] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News. *arXiv preprint arXiv:1703.06959* (2017).
[46] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *WWW*.
[47] Kate Starbird and Leysia Palen. 2010. Pass It On?: Retweeting in Mass Emergency. In *Proceedings of the 7th International ISCRAM Conference*.
[48] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107* (2017).
[49] Nguyen Vo, Kyumin Lee, Cheng Cao, Thanh Tran, and Hongkyu Choi. 2017. Revealing and detecting malicious retweeter groups. In *ASONAM*.
[50] Nguyen Vo, Kyumin Lee, and Thanh Tran. 2017. MRAttractor: Detecting communities from large-scale graphs. In *IEEE BigData*. IEEE, 797–806.
[51] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *KDD*.
[52] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In *KDD*.
[53] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *ICDE*.
[54] Liang Wu and Huan Liu. 2018. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *WSDM*.
[55] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In *WSDM*.
[56] Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. 2013. Recommending Fresh URLs Using Twitter Lists.. In *ICWSM*.
[57] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *WWW*.
[58] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *WWW*.
[59] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11, 3 (2016), e0150989.