Collective Classification for Social Media Credibility Estimation

Kyle O'Brien[1], Olga Simek[2], Frederick R. Waugh[2]
[1] *Amazon, Inc., Cambridge, MA, United States*
[2] *MIT Lincoln Laboratory, Lexington, MA, United States*

Abstract

We introduce a novel extension of the iterative classification algorithm to heterogeneous graphs and apply it to estimate credibility in social media. Given a heterogeneous graph of events, users, and websites derived from social media posts, and given prior knowledge of the credibility of a subset of graph nodes, the approach iteratively converges to a set of classifiers that estimate credibility of the remaining nodes. To measure the performance of this approach, we train on a set of manually labeled events extracted from a corpus of Twitter data and calculate the resulting receiver operating characteristic (ROC) curves. We show that collective classification outperforms independent classification approaches, implying that graph dependencies are crucial to estimating credibility in social media.

1. Introduction

1.1. Credibility in Social Media

Social media are an increasingly important pathway for real-time dissemination of information about breaking news. For unfolding events as diverse as natural disasters [1], terror attacks [2], and sociopolitical protests [3], social media platforms such as Twitter and Facebook can serve as a vital communications infrastructure providing time-critical information. Unfortunately, the speed and effectiveness with which information propagates through social media have also attracted users seeking to spread disinformation, rumors, and propaganda [4, 5].

Recently, Twitter, Facebook, and Google have acknowledged the susceptibility of their platforms to such disinformation [6, 7] and have taken steps to fight harmful use of their services. Additionally, independent fact-checking organizations—including Snopes.com, Politifact, and FactCheck.org—have arisen that seek to detect disinformation. However, their manual approach to fact checking cannot scale to address the enormous data volumes generated by large social networks. Similarly, efforts at crowd-sourced fact checking, such as Web of Trust and 4facts.org, cannot keep pace with the rate of online social media activity.

1.2. Automated Credibility Assessment

Machine learning techniques offer promise for assessing social media content and source credibility at scale. Conventional machine learning models that assume data to be independent and identically distributed (i.i.d.) have been used for credibility estimation [8] and form the basis of TweetCred, an online tool for automated credibility estimation for Twitter [9]. However, i.i.d. approaches neglect valuable predictive information that arises from interactions among social media users. For example, two social media users who post a link to the same news website are more likely to share other properties, such as political beliefs, demographics, community membership, and common friends [10].

A central hypothesis of this paper is that this tendency of linked entities to share properties, known as homophily, extends to credibility in social networks. To exploit this effect, we propose an automated, end-to-end system for social media credibility assessment, depicted in Figure 1, that consists of modules for acquiring social media posts, detecting clusters of posts that correspond to specific events, constructing a relationship graph linking important entities, and finally estimating a credibility score for all graph entities.

While all four of these modules present technical challenges, this paper focuses on algorithms for the credibility estimation module. In particular, we describe a novel collective classification algorithm [11-14] that exploits correlations in a heterogenous graph of users, events, and websites in order to improve on traditional machine learning approaches for data that are not i.i.d. Figure 2 shows what a portion of the graph looks like for an example noncredible event in our data set ("Iraq's army has shot down two British planes as they were carrying weapons for the ISIS"). The results presented in this paper demonstrate that, by exploiting homophilic interactions in this network, our algorithm yields improved performance over independent classification approaches.

URI: https://hdl.handle.net/10125/59662
ISBN: 978-0-9981331-2-6

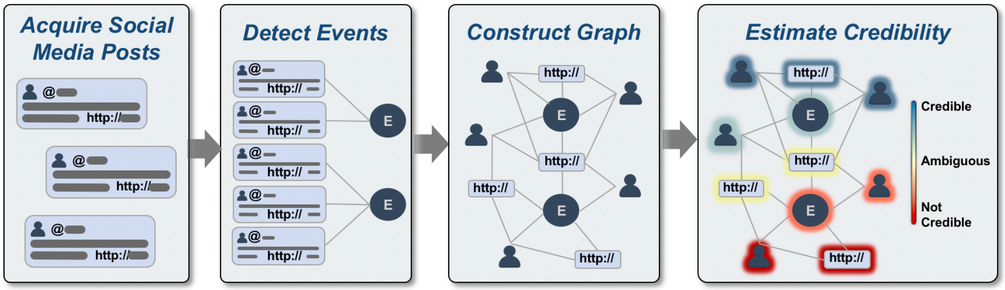


Figure 1. Proposed system for automated social media credibility assessment, consisting of modules for acquiring social media posts, detecting events, constructing a relationship graph, and estimating credibility of graph entities. This paper focuses on the credibility estimation module. In the figure, users are indicated by a person icon, websites by http://, and events by E. Color shading in the estimation module indicates notional output of a credibility estimation algorithm.

Our approach can also leverage existing databases of untrustworthy users and websites for training, allowing the system to evolve as new trustworthiness labels are provided. We evaluate the approach by extracting credible and noncredible events from a large corpus of Twitter data and quantifying how well the system classifies those events as well as the users and websites associated with them.

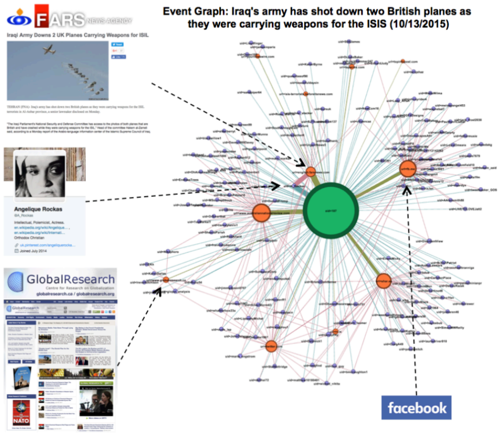

Event Graph: Iraq's army has shot down two British planes as they were carrying weapons for the ISIS (10/13/2015)
FARS
Iraqi Army Downs 2 UK Planes Carrying Weapons for ISIL
Angelique Rockas
GlobalResearch
facebook

Figure 2. Example graph for the event "Iraq's army has shot down two British planes as they were carrying weapons for the ISIS," showing single event node (green) with associated user nodes (blue) and website nodes (orange).

1.3. Related Work

Collective Classification. Early approaches to predicting credibility of events in social media used supervised classification models with i.i.d. assumptions [8]. However, social media data form rich networks that encode important non-i.i.d. dependencies. Collective classification techniques can improve predictions by learning not only from local features (those specific to a given node) but also from relational features (which capture behavior of neighboring nodes).

Notable collective classification approaches include loopy belief propagation (LBP), mean field relaxation labeling (MFRL), Gibbs sampling (GS), and the iterative classification algorithm (ICA) [11-13]. LBP and MFRL are global methods, defining a single objective function and searching for its optimum. In contrast, GS and ICA are both local methods that use classifiers in an iterative fashion to converge to a self-consistent solution.

A major challenge to applying these algorithms to social media data is the heterogeneity of the associated entity graph, which results in widely varying local and relational feature statistics, so that a single estimator cannot be expected to generalize well across all node types. Hybrid classifiers, which use one classifier for local features and another for relational features [14], do not address the need for separate classifiers for different node types.

Another challenge to classification in networked data is the sparsity of typical relational training sets. Some approaches assume a fully truthed training graph, impractical for large social media datasets. Other approaches use bootstrapping methods for training relational classifiers that produce predicted labels to augment the known ones [11, 12], the approach taken in this paper.

The optimal choice of graph entities and relationships for effective event credibility prediction is an open-ended problem. This paper shows how constructing a graph of events, users, and linked websites allows us to leverage credibility labels from existing fact-checking websites and inputs from users to train and continually improve the algorithm. Incorporating other entities into the graph may improve performance; see Section 6 for further discussion.

Credibility Estimation and Rumor Detection. A number of approaches have been reported for estimating credibility or for detecting rumors or false information in social media [4, 8, 15-19]. This paper builds on many of these. For example, we leverage key features of social media posts that are shown by Castillo *et al.* [8] to be associated with credibility, and diffusion approaches [4, 17] resemble our collective classification algorithm in some ways. However, these approaches either assume i.i.d. data [8, 19] or limit themselves to estimating the credibility only of the message or event being reported [4, 8, 16, 17], but not of users, websites, or other entities. In contrast, this paper incorporates non-i.i.d. correlations and addresses the credibility not only of messages but also of associated social media users and websites. Other, more recent approaches that analyze the dynamics of rumor propagation [16] demonstrate intriguing results that may be incorporated into our approach in the future (see Section 6 for further discussion).

In addition, significant work has been done on estimating credibility in online reviews and other electronic word-of-mouth sources [20-22]. Many of these treat reviews independently, using regression [20] or classifiers [21] to identify features associated with review content and reviewer behavior that correlate with review credibility. The results described in this paper indicate that explicitly incorporating network effects improves credibility estimation compared to approaches that assume independence for the case of event reporting in social media.

Finally, end-to-end credibility estimation systems like that of Figure 1 have been proposed [23, 24] and demonstrated [9]. These differ from our approach in that they use independent classifiers [9, 23] or focus on assessing content credibility only [24].

2. Preliminaries

In this section, we discuss background information and terminology used throughout the remainder of the paper.

2.1. Classification in Networked Data

First, we introduce notation and state the general networked classification problem [11-13]. Assume we are given a set of N vertices $v = \{v_1, \ldots, v_N\}$ where each vertex v_i has a corresponding feature vector $X_i \in \mathbb{R}^M$, where M is the feature space dimension, and where some vertices also have a label $Y_i \in \mathbb{N}$. Vertices are connected to one another by a set of edges $e \subseteq v \times v$ that encode relational dependencies, forming a graph $G = (v, e)$. The neighborhood function $\mathcal{N}_i$ defines the adjacent vertices of v_i such that $\mathcal{N}_i \subseteq v \backslash v_i$.

Given a graph G, features X for all vertices, and known labels $Y^K \subset Y$ for a subset $v^K \subset v$ of vertices, the collective classification task is to infer the unknown labels $Y^U \subset Y \backslash Y^K$. In contrast, independent classification assumes no dependencies and thus no graph, and the goal is to infer Y^U given training set $\{X^K, Y^K\}$ and features X^U.

Depending on the particular problem domain, each vertex represents some meaningful entity in the data and the graph topology is inferred from observed network structure that encodes relational dependencies. For example, consider predicting the topics of webpages using both their content and the graph that encodes the hyperlinks among them. Because a given webpage will tend to have hyperlinks to webpages about similar topics, including the graph adds greater predictive information to the classification problem compared to using webpage content alone.

2.2. Iterative Classification

We chose ICA for the work presented in this paper for several reasons. First, it has been shown that in many cases global methods provide little if any performance advantage over local methods like ICA while adding a significant computational burden that scales poorly with network size [11, 12]. Second, the simplicity and flexibility of ICA enabled its straightforward extension to heterogeneous graphs. Finally, exploratory work using an algorithm developed specifically for social media analysis [17] failed to show improvement over independent classification approaches for our data set.

ICA predicts vertex labels in a graph using a classifier *f*, which uses features derived from attributes and labels in the neighborhood $\mathcal{N}_i$ of vertex v_i to estimate the best class label Y_i. We iterate through each vertex and update node predictions until assignments stabilize or until a maximum number of iterations is reached. For a given vertex v_i, the input to *f* is a feature vector $X_i = \left[X_i^{loc}, X_i^{rel}\right]^{\mathrm{T}}$ that combines local features X_i^{loc}, derived from the attributes of v_i, and relational features X_i^{rel}, derived from the neighbors $\mathcal{N}_i$ of v_i. The relational features may be any function of the attributes and labels of $\mathcal{N}_i$; aggregate statistics are commonly used, such as the count, mean, standard deviation, or maximum value of the neighboring labels. On the first iteration, labels Y^U are unknown and therefore some X_i^{rel} will have undefined values, so we first bootstrap labels by predicting Y^U using local features X_i^{loc} only.

3. Event Credibility Algorithm

This section introduces our algorithm for event credibility estimation in social media. The intuition behind our approach is that social media activity forms rich networks that can propagate important information for assessing credibility of events. By choosing a suitable network structure, we can improve event credibility estimation over non-networked approaches. Specifically, we expect users who seek to spread misinformation to contribute to false or hoax events, so if we know the credibility of one event, then we can infer credibility of other events involving the same users. In turn, those users tend to promote websites whose content reinforces misinformation and news bias, so if we know the credibility of a website, then we can infer credibility of events linked to those websites.

In this section, we formalize these intuitions by describing how to construct a heterogeneous network of users, events, and websites and how to extend ICA to enable inference on that network. This novel extension of ICA to heterogeneous graphs is the key advance that enables our approach to outperform independent classification in estimating social media credibility.

3.1. Event Extraction

Event extraction is the process of finding social media posts written about real or hoax events. Given a set D of social media posts $d \in D$, we denote $E_i \in D$ to be the subset of posts that discuss event *i*. E_i can be described by a function $Q_i(d)$ which equals 1 for posts about event *i* and 0 otherwise: $E_i = \{d \in D | Q_i(d) = 1\}$. Event extraction, then, is the determination of $Q_i(d)$ for all i. In practice, posts are typically stored in an indexed database, and $Q_i(d)$ represents the database query that retrieves E_i.

While automated approaches exist for event detection [25], we considered their performance to be insufficient for producing the high-quality inputs we needed for credibility estimation. As a result, we used a semi-automated approach in which queries $Q_i(d)$ are interactively constructed and refined using boolean combinations of distinguishing keywords. For example, posts associated with the event shown in Figure 2 were retrieved using the initial query $Q =$ `{"(iraqi army) AND shot AND (british AND (plane OR jet)) AND (weapons OR arms OR guns) AND (isis or isil or daesh)"},` which was then refined based on the contents of the returned posts. We intend to incorporate automated event extraction into future versions of this system (see Section 6).

3.2. User and Website Extraction

Event extraction yields a set of social media posts $D' = \{d \mid d \in \bigcup E_i\}$ related to all events. This section describes how user and websites entities are extracted from D'.

Like events, users and websites comprise sets of social media posts extracted from D':

- U_j is the set of posts authored by user j: $U_j = \{d \in D' | \text{ user } j \text{ authored } d\}$;
- W_k is the set of posts that contain website *k*: $W_k = \{d \in D' \mid \text{website } k \text{ appears in } d\}$.

Websites are identified using their uniform resource locator (URL). URLs are processed by expanding any hyperlinks shortened with a service such as *bitly.com* or *tinyurl.com* and by stripping to the top-level domain. For example, the URL *http://website.com/news-article-title* is stripped to *http://website.com*. This is done to limit the granularity of hyperlinks to gather better supporting statistics for relational credibility features. While a particular news article webpage typically refers to a single event, its top-level domain is likely to be associated with multiple events.

3.3. Graph Construction

The graph is constructed using the following rules for creating edges:

- e_{E_i,U_j}: An edge is created between event E_i and user U_j for each post in E_i authored by U_j;
- e_{U_j,W_k}: An edge is created between user U_j and website W_k for each post authored by U_j that contains W_k;
- e_{E_i,W_k}: An edge is drawn between event E_i and website W_k for each post in E_i that contains W_k.

This process generates a heterogeneous, undirected, weighted graph G that connects events, users, and websites. Using weighted rather than binary edges allows greater influence between strongly connected entities. For example, the credibility of an event connected to both credible and noncredible users is the weighted average of their contributions.

Other rule sets leading to other graph structures are possible and may in fact lead to better performance; see Section 6 for further discussion.

3.4. Entity Credibility Database

After graph construction, node labels are required to train the classifier models. We compiled a database of credible and noncredible websites and users by combining results from publicly available fact checking services—such as Web of Trust, 4facts.org, Politifact.com, and Factcheck.org—as well as from our own manual credibility analysis. This database is also a mechanism where users of the system can collectively contribute labels to improve its accuracy.

3.5. Iterative Classification

Once the graph is constructed and the entity credibility database is populated, we proceed with iterative classification. Because local feature distributions around a given node can vary strongly depending on whether that node is an event, user, or website, we use different estimators for each node type. The graph schema implies six different estimator models: one each for event, user, and website node types for learning labels Y^U in the initial bootstrapping phase, and one each for event, user, and website node types in the iterative classification phase. The best-performing model is chosen to maximize the final estimates.

4. Experimental Setup

In this section we describe experiments performed on a corpus of Twitter data using the credibility estimation techniques described in the previous section.

4.1 Data

We started with a database D of approximately 50 million tweets about current events in the Middle East, acquired over a continuous, 90-day interval. We manually extracted over 100 events from D and assessed their credibility by researching online reporting, including news articles, blogs, and fact-checking sites. In contrast to others who have used a Likert scale to account for ambiguity [19], we tagged events as either credible or noncredible, using the following approach. Events reported in multiple mainstream news outlets were considered to be credible. Events either explicitly discredited by fact-checking sites or reported only by outlets known for propaganda or false information were considered noncredible. A small number of events not in either category were simply discarded.

Using this process, we generated 68 credible events and 39 noncredible events. Database queries Q_i, $i = 1, \ldots, 107$, were then created to extract tweets associated with each event from the database. In total, 356,087 event-related tweets were extracted to form D'. From those, we extracted 63,944 unique users and 3,894 unique websites and built a graph using the procedure described in Section 3.

4.2 Features

Next, we identified local and relational features for building classifiers. Local features were extracted from the tweets and tweet metadata associated with each event, user, and website node in the graph. The particular local features used, shown in Table 1, combine features reported elsewhere to be discriminative [8] with some that we found to be effective. Local features are computed over all tweets associated with an entity: for example, the feature "Average age of tweets" is the average age of all tweets associated with a given event, user, or website.

Table 1. Local features for independent and collective classifiers.

Local Features	
Type	**Description**
Tweet	Average age of tweets
	Average number of characters per tweet
	Average number of words per tweet
	Fraction of tweets containing question mark
	Fraction of tweets containing exclamation mark
	Fraction of tweets containing emoji
	Fraction of tweets containing first person pronoun
	Fraction of tweets containing user mention
	Fraction of tweets that are retweets
	Fraction of tweets by most frequently-occurring author
User	Average age of user accounts
	Average number of followers per user account
	Average number of friends per user account
	Fraction of distinct users
URL	Fraction of tweets having a URL
	Number of distinct URLs

In addition to local features, we computed relational features that capture dependencies among graph nodes.

As shown in Table 2, we chose four relational features for each node type to form X_i^{rel}: the fraction of credible neighbors and the average neighbor credibility value after grouping neighbors by node type. Local and relational features were then concatenated to form feature vectors $X_i = \left[X_i^{loc},\ X_i^{rel}\right]^{\mathrm{T}}$.

Table 2. Relational features for collective classifiers.

	Relational Features
Type	**Description**
Event	Fraction of credible user neighbors
	Average credibility of user neighbors
	Fraction of credible URL neighbors
	Average credibility of URL neighbors
User	Fraction of credible URL neighbors
	Average credibility of URL neighbors
	Fraction of credible user neighbors
	Average credibility of user neighbors
URL	Fraction of credible user neighbors
	Average credibility of user neighbors
	Fraction of credible event neighbors
	Average credibility of event neighbors

4.3 Training

In the label bootstrapping phase, unknown labels Y^U were initialized using independent classifiers trained on local features only. We found that logistic regression (LR) with balanced class weights worked well under 5-fold cross validation.

Once initial estimates for Y^U are obtained, relational classifiers were trained using local and relational features. We found the following models to work well for the relational classification task: a decision tree (DT) with maximum depth of 4 for events, and a LR model with balanced class weights for users and URLs.

5. Credibility Estimation Results

Credibility estimation results are reported in this section for event, user, and URL nodes. To avoid overfitting, we used stratified cross validation. We split all event, user, and URL nodes independently into 5 stratified folds, with the proportions of labeled and unlabeled nodes equal across folds to mitigate class imbalance. For each fold, we trained classifiers on the training fold and tested them on the test fold.

Finally, we produced receiver operating characteristic (ROC) curves for each fold and averaged these to report a final ROC curve. We computed both the area under the curve (AUC) and overall classification accuracy as performance measures. This evaluation was performed for both independent classification, which ignores the graph and relational classifiers, and for ICA, which includes them.

5.1 Event Credibility Performance

Figure 2 compares event credibility ROC curves for independent and collective classifiers. The shaded region around each curve shows ±1 standard deviation of the stratified cross validation folds to indicate the range of performance over the whole set of folds. We achieve an overall average accuracy of 95.2% for predicting event credibility. Iterative classification improves the AUC by 7.3% and the accuracy by 22.9% over independent classification methods.

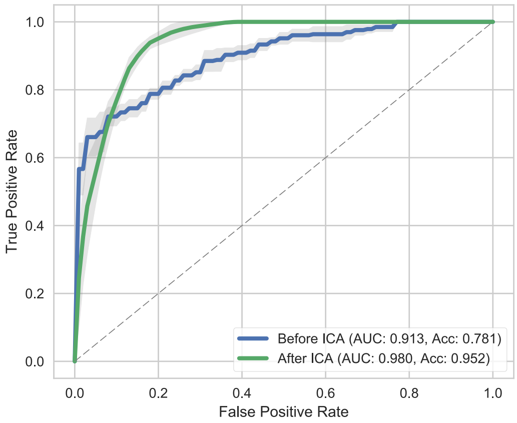
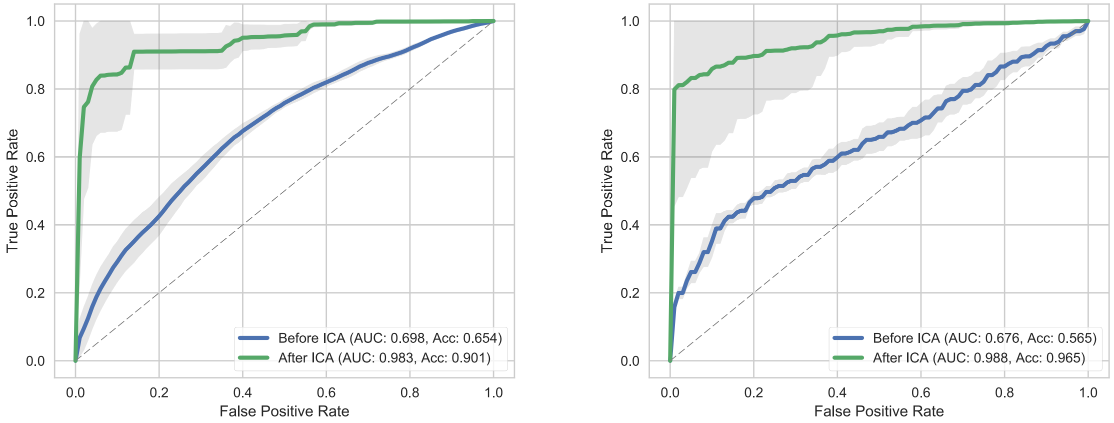


Figure 2. ROC curves for event credibility estimation, comparing performance of independent classification (blue) and ICA (green). Shaded area shows ±1 standard deviation of 5-fold cross validated results. ICA improves both AUC and accuracy.

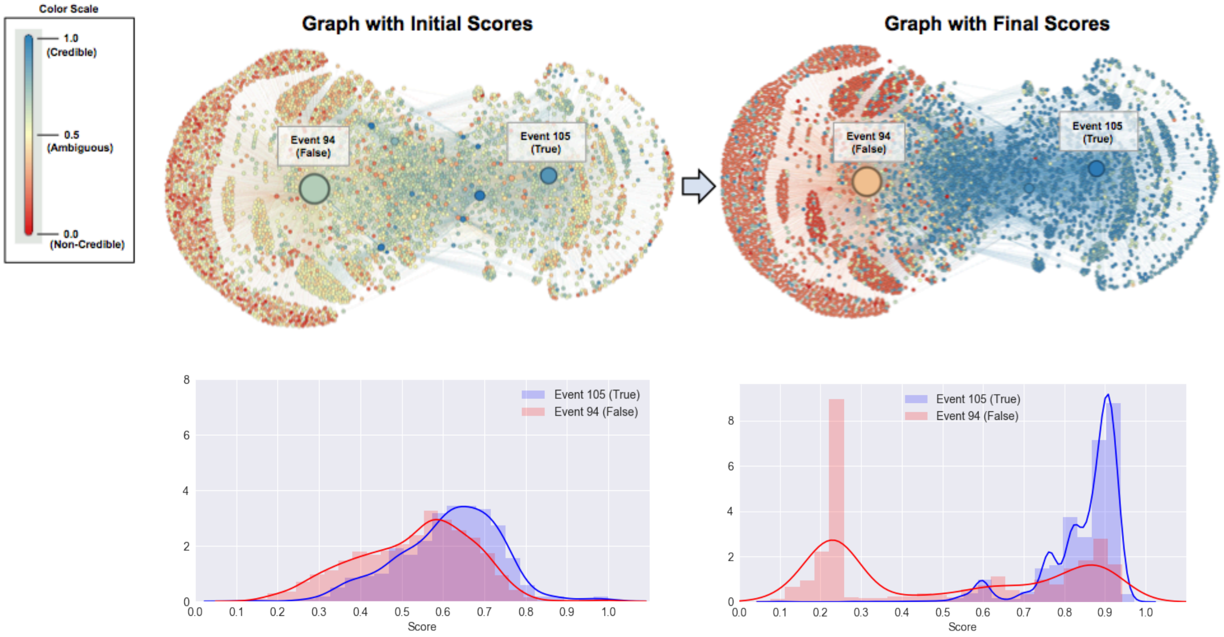
5.2 User and URL Credibility Performance

ROC curves for users and URLs appear in Figure 3, showing that iterative classification improves the AUC and accuracy for these entity types as well. We achieve 90.1% overall accuracy for users and 96.5% overall accuracy for URLs. User AUC improves by 40.8% and user accuracy improves by 37.8% compared to independent classification. URL AUC improves by 46.2% and URL accuracy improves by 70.8% over independent classification.



(a) (b)

Figure 3. ROC curves for (a) user and (b) URL credibility estimation, comparing independent (blue) and iterative (green) classifiers. Shaded area shows ±1 standard deviation of 5-fold cross validated results.



(a) (b)

Figure 4. Example results for (a) independent and (b) ICA classifiers for two linked events, one credible and one noncredible, showing how ICA boosts credibility estimates for credible events and their associated entities. Top shows event credibility graphs colored by credibility score; bottom shows histograms of credibility scores. The credible event is "ISIS militants shoot dead 70 members of Sunni Albu Nimr tribe in Iraq" (event 105); the noncredible event is "Turkish AKP Party has sent weapons to ISIS terrorist organizations" (event 94).

5.3 Performance Example

Figure 4 shows how to visualize the effect of applying ICA to independent classifier results for a specific case of two linked events, one credible and one noncredible, demonstrating that ICA exploits relationships among events, users, and URLs to provide additional discrimination compared to independent classification.

6. Conclusions and Future Work

In this paper we have proposed an automated, end-to-end system that uses collective classification techniques to exploit homophily in a network of social media events, users, and websites in order to estimate credibility of these entities. We have demonstrated techniques for producing a graph of relationships among event, user, and website nodes and for extending collective classification approaches to heterogeneous node types. We have shown that credibility estimation performance is enhanced when leveraging this graph because labels of entities are correlated with the labels of their neighbors.

We anticipate future work in the following areas to further improve system performance.

Network structure. The graph structure used in this paper—consisting of users, events, and websites linked through the rules listed in Section 3.3—is one of a number of possible structures. Investigating these structures to determine which provides best estimation performance is a topic of future research. One possibility is to add links to the node set used here, for example by linking users who follow, message, or retweet each other. Another possibility is to add other node types, such as tweets (which are treated in bulk in this paper), offering the possibility of estimating credibility at the individual tweet level. Our initial work in this area has yet to yield reliable credibility estimates for individual tweets.

Network dynamics. Our approach treats the graph of users, events, and websites as static and estimates credibility based on time-independent properties of these nodes and their links. Other researchers [16, 26] have shown that tweet dynamics and propagation provide important information about credibility. We intend to explore the utility of time-dependent features in future work. We also plan to assess event credibility timeliness by evaluating how much data is needed before confident credibility estimates can be made.

Event detection. The work presented in this paper addresses only one module of the proposed end-to-end automated system depicted in Figure 1. In particular, as discussed in Section 4.1, we used a semi-automated process for event detection. In future work, we plan to replace this process with a fully automated event detection module [25]. As a first step, we have investigated improving the quality of tweets used in prediction by first automatically filtering on whether or not the tweet represents an assertion rather than an opinion, question, or other construct. In preliminary work, we have shown that deep learning approaches utilizing long-short term memory recurrent neural networks on tweet text can achieve human-level performance for predicting assertions.

This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Air Force.

References

[1] Neubig, G.; Matsubayashi, Y.; Hagiwara, M.; and Murakami, K. Safety information mining: What can NLP do in a disaster? In *Proc. 5th International Joint Conference on Natural Language Processing* (2011).

[2] Burnap, P.; Williams, M. L.; Sloan, L.; Rana, O.; Housley, W.; Edwards, A.; Knight, V.; Procter, R.; and Voss, A. Tweeting the terror: modeling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining* 4, 943–965 (2014).

[3] Gonzalez-Bailon, S.; Borge-Holthoefer, J.; and Moreno, Y. Broadcasters and hidden influentials in online protest diffusion. *American Behavioral Scientist* 57(7):943–965 (2013).

[4] Lee, J.; Agrawal, M.; and Rao, H. R. Message diffusion through social network service: the case of rumor and non-rumor related tweets during Boston bombing 2013. *Information Systems Frontiers* 17, 997–1005 (2015).

[5] Paul, C., and Matthews, M. The Russian firehose of falsehood propaganda model. Technical report, RAND Corporation (2016).

[6] Weedon, J.; Nuland, W.; and Stamos, A. Information operations and Facebook. Technical report, Facebook, Inc. (2017).

[7] Breland, A. 2017. Social media fights back against fake news. *The Hill.com*, 27 May 2017.

[8] Castillo, C.; Mendoza, M.; and Poblete, B. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, New York, NY, USA: ACM, 675–684 (2011).

[9] Gupta, A.; Kumaraguru, P.; Castillo, C.; and Meier, P. TweetCred: Real-time credibility assessment of content on Twitter. In L.M. Aiello and D. McFarland, eds., *SocInfo*

2014, LNCS 8851, 228-243. Switzerland: Springer International Publishing (2014).

[10] McPherson, M.; Smith-Lovin, L.; and Cook, J. M. Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* 27, 415–44 (2001).

[11] Sen, P.; Namata, G.; Bilgic, M.; and Getoor, L. Collective classification. Boston, MA: Springer US, 189–193 (2010).

[12] Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Gallagher, B.; and Eliassi-Rad, T. Collective classification in network data. *AI Magazine* 29(3), 93–106 (2008).

[13] Neville, J.; and Jensen, D. D. Iterative classification in relational data. In Getoor, L., and Jensen, D. D., eds., *Proc. Workshop on Learning Statistical Models from Relational Data, Seventeenth National Conference on Artificial Intelligence*, 42–49. Austin, TX: AAAI Press, Menlo Park, CA (2000).

[14] Cataltepe, Z.; Sonmez, A.; Baglioglu, K.; and Erzan, A. Collective classification using heterogeneous classifiers. Berlin, Heidelberg: Springer, 155–169 (2011).

[15] Zubiaga, A.; Aker, A.; Bontcheva, K.; Liakata, M.; and Procter, R. Detection and resolution of rumors in social media: a survey. *ACM Computing Surveys* 51(2) (2016).

[16] Vosoughi, S.; Roy, D.; and Aral, S. The spread of true and false news online. *Science* 359 (6380), 1146–1151 (2018).

[17] Gupta, M.; Xhao, P.; and Han, J. Evaluating event credibility on Twitter. *Proc. 2012 SIAM International Conference on Data Mining* (2012).

[18] Algarni, A.; Xu, Y.; and Chan, T. An empirical study on the susceptibility to social engineering in social networking sites: the case of Facebook. *European Journal of Information Systems* 26(6), 661-87 (2017).

[19] O'Donovan, J.; Kang, B.; Meyer, G.; Hollerer; T.; and Adalfi, A. Credibility in context: an analysis of feature distributions in Twitter. *2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, 293-301 (2012).

[20] Banerjee, Shankhadeep; Samadrita Bhattacharyya; and Indranil Bose. Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business. *Decision Support Systems* 96, 17 (2017).

[21] Zhang, Dongsong, *et al.* What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems* 33 (2), 456 (2016).

[22] Fang, Y.-H. Beyond the credibility of electronic word of mouth: exploring eWOM adoption on social networking sites from affective and curiosity perspectives. *International Journal of Electronic Commerce* 18 (3), 67 (2014).

[23] Liu, X., *et al.* Towards a highly effective and robust web credibility evaluation system. *Decision Support Systems* 79, 99 (2015).

[24] Fuller, C. M.; D. P. Biros; and R. L. Wilson. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems* 46 (3) (2009): 695.

[25] Atefeh, F.; and Khreich, W. A survey of techniques for event detection in Twitter. *Computational Intelligence* 31(1), 132–164 (2015).

[26] Metaxas, P.; and Mustafaraj, E. Trails of trustworthiness in real-time streams. *Design, Influence and Social Technologies Workshop of CSCW12* (2012).