Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes

Srijan Kumar*
University of Maryland
srijan@cs.umd.edu

Robert West
Stanford University
west@cs.stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Wikipedia is a major source of information for many people. However, false information on Wikipedia raises concerns about its credibility. One way in which false information may be presented on Wikipedia is in the form of hoax articles, *i.e.*, articles containing fabricated facts about nonexistent entities or events. In this paper we study false information on Wikipedia by focusing on the hoax articles that have been created throughout its history. We make several contributions. First, we assess the real-world impact of hoax articles by measuring how long they survive before being debunked, how many pageviews they receive, and how heavily they are referred to by documents on the Web. We find that, while most hoaxes are detected quickly and have little impact on Wikipedia, a small number of hoaxes survive long and are well cited across the Web. Second, we characterize the nature of successful hoaxes by comparing them to legitimate articles and to failed hoaxes that were discovered shortly after being created. We find characteristic differences in terms of article structure and content, embeddedness into the rest of Wikipedia, and features of the editor who created the hoax. Third, we successfully apply our findings to address a series of classification tasks, most notably to determine whether a given article is a hoax. And finally, we describe and evaluate a task involving humans distinguishing hoaxes from non-hoaxes. We find that humans are not good at solving this task and that our automated classifier outperforms them by a big margin.

1. INTRODUCTION

The Web is a space for all, where, in principle, everybody can read, and everybody can publish and share, information. Thus, knowledge can be transmitted at a speed and breadth unprecedented in human history, which has had tremendous positive effects on the lives of billions of people. But there is also a dark side to the unreigned proliferation of information over the Web: it has become a breeding ground for false information [6, 7, 12, 15, 19, 43].

The reasons for communicating false information vary widely: on the one extreme, *misinformation* is conveyed in the honest but mistaken belief that the relayed incorrect facts are true; on the other extreme, *disinformation* denotes false facts that are conceived in order to deliberately deceive or betray an audience [11, 17]. A third class of false information has been called *bullshit,* where the agent's primary purpose is not to mislead an audience into believing false facts, but rather to "convey a certain impression of himself" [14].

All these types of false information are abundant on the Web, and regardless of whether a fact is fabricated or misrepresented on purpose or not, the effects it has on people's lives may be detrimental and even fatal, as in the case of medical lies [16, 20, 22, 30].

Hoaxes. This paper focuses on a specific kind of disinformation, namely *hoaxes*. Wikipedia defines a hoax as "a deliberately fabricated falsehood made to masquerade as truth." The Oxford English Dictionary adds another aspect by defining a hoax as "a *humorous* or mischievous deception" (italics ours).

We study hoaxes in the context of Wikipedia, for which there are two good reasons: first, anyone can insert information into Wikipedia by creating and editing articles; and second, as the world's largest encyclopedia and one of the most visited sites on the Web, Wikipedia is a major source of information for many people. In other words: Wikipedia has the potential to both attract and spread false information in general, and hoaxes in particular.

The impact of some Wikipedia hoaxes has been considerable, and anecdotes are aplenty. The hoax article about a fake language called "Balboa Creole French", supposed to be spoken on Balboa Island in California, is reported to have resulted in "people coming to [...] Balboa Island to study this imaginary language" [38]. Some hoaxes have made it into books, as in the case of the alleged (but fake) Aboriginal Australian god "Jar'Edo Wens", who inspired a character's name in a science fiction book [10] and has been listed as a real god in at least one nonfiction book [24], all before it came to light in March 2015 that the article was a hoax. Another hoax ("Bicholim conflict") was so elaborate that it was officially awarded "good article" status and maintained it for half a decade, before finally being debunked in 2012 [27].

The list of extreme cases could be continued, and the popular press has covered such incidents widely. What is less available, however, is a more general understanding of Wikipedia hoaxes that goes beyond such cherry-picked examples.

Our contributions: impact, characteristics, and detection of Wikipedia hoaxes. This paper takes a broad perspective by starting from the set of all hoax articles ever created on Wikipedia and illuminating them from several angles. We study over 20,000 hoax articles, identified by the fact that they were explicitly flagged as potential hoaxes by a Wikipedia editor at some point and deleted after a discussion among editors who concluded that the article was

*Research done partly during a visit at Stanford University.

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883085.

indeed a hoax. Some articles are acquitted as a consequence of that discussion, and we study those as well.

When answering a question on the Q&A site Quora regarding the aforementioned hoax that had been labeled as a "good article", Wikipedia founder Jimmy Wales wrote that "[t]he worst hoaxes are those which (a) last for a long time, (b) receive significant traffic and (c) are relied upon by credible news media" [33]. Inspired by this assessment, our first set of questions aims to understand how impactful (and hence detrimental, by Wales's reasoning) typical Wikipedia hoaxes are by quantifying (a) how long they last, (b) how much traffic they receive, and (c) how heavily they are cited on the Web. We find that most hoaxes have negligible impact along all of these three dimensions, but that a small fraction receives significant attention: 1% of hoaxes are viewed over 100 times per day on average before being uncovered.

In the second main part of the paper, our goal is to delineate typical characteristics of hoaxes by comparing them to legitimate articles. We also study how successful (*i.e.*, long-lived and frequently viewed) hoaxes compare to failed ones, and why some truthful articles are mistakenly labeled as hoaxes by Wikipedia editors. In a nutshell, we find that on average successful hoaxes are nearly twice as long as legitimate articles, but that they look less like typical Wikipedia articles in terms of the templates, infoboxes, and inter-article links they contain. Further, we find that the "wiki-likeness" of legitimate articles wrongly flagged as hoaxes is even lower than that of actual hoaxes, which suggests that administrators put a lot of weight on these superficial features when assessing the veracity of an article.

The importance of the above features is intuitive, since they are so salient, but in our analysis we find that less immediately available features are even more telling. For instance, new articles about real concepts are often created because there was a need for them, reflected in the fact that the concept is mentioned in many other articles before the new article is created. Hoaxes, on the contrary, are mentioned much less frequently before creation—they are about nonexistent concepts, after all—but interestingly, many hoaxes still receive some mentions before being created. We observe that such mentions tend to be inserted shortly before the hoax is created, and by anonymous users who may well be the hoaxsters themselves acting *incognito*.

The creator's history of contributions made to Wikipedia before a new article is created is a further major distinguishing factor between different types of articles: most legitimate articles are added by established users with many prior edits, whereas hoaxes tend to be created by users who register specifically for that purpose.

Our third contribution consists of the application of these findings by building machine-learned classifiers for a variety of tasks revolving around hoaxes, such as deciding whether a given article is a hoax or not. We obtain good performance; *e.g.*, on a balanced dataset, where guessing would yield an accuracy of 50%, we achieve 91%. To put our research into practice, we finally find hoaxes that have not been discovered before by running our classifier on Wikipedia's entire revision history.

Finally, we aim to assess how good humans are at telling apart hoaxes from legitimate articles in a typical reading situation, where users do not explicitly fact-check the article by using a search engine, following up on references, etc. To this end, we design and run an experiment involving human raters who are shown pairs consisting of one hoax and one non-hoax and asked to decide which one is the hoax by just inspecting the articles without searching the Web or following links. Human accuracy on this task is only 66% and is handily surpassed by our classifier, which achieves 86% on the same test set. The reason is that humans are biased to believe that well-formatted articles are legitimate and real, whereas it is easy for our classifier to see through the markup glitter by also considering features computed from other articles (such as the number of mentions the article in question receives) as well as the creator's edit history.

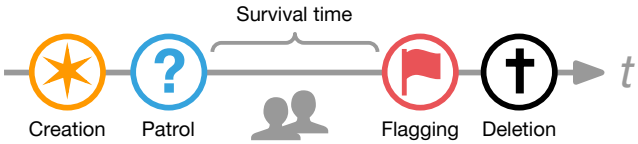


Figure 1: Life cycle of a Wikipedia hoax article. After the article is created, it passes through a human verification process called patrol. The article survives until it is flagged as a hoax and eventually removed from Wikipedia.

The remainder of this paper is structured as follows. Sec. 2 outlines the life cycle Wikipedia hoaxes go through from creation to deletion. In Sec. 3, 4, and 5 we discuss the impact, characteristics, and automated detection of hoaxes, respectively. The experiment with human subjects is covered in Sec. 6. Related work is summarized in Sec. 7; and Sec. 8 concludes the paper.

2. DATA: WIKIPEDIA HOAXES

The Wikipedia community guidelines define a hoax as "an attempt to trick an audience into believing that something false is real", and therefore consider it "simply a more obscure, less obvious form of vandalism" [39].

A distinction must be made between *hoax articles* and *hoax facts*. The former are entire articles about nonexistent people, entities, events, etc., such as the fake Balboa Creole French language mentioned in the introduction.[1] The latter are false facts about existing entities, such as the unfounded and false claim that American journalist John Seigenthaler "was thought to have been directly involved in the Kennedy assassinations" [40].

Finding hoax facts is technically difficult, as Wikipedia provides no means of tagging precisely one fact embedded into a mostly correct article as false. However, in order to find hoax articles, it suffices to look for articles that were flagged as such at some point. Hence we focus on hoax articles in this paper.

To describe the mechanism by which hoax articles are flagged, we need to consider Wikipedia's page creation process (schematized in Fig. 1). Since January 2006 the privilege of creating new articles has been limited to logged-in users (*i.e.*, we know for each new article who created it). Once the article has been created, it appears on a special page that is monitored by trusted, verified Wikipedians who attempt to determine the truthfulness of the new article and either mark it as legitimate or flag it as suspicious by pasting a template[2] into the wiki markup text of the article.

This so-called patrolling process (introduced in November 2007) works very promptly: we find that 80% of all new articles are patrolled within an hour of creation, and 95% within a day. This way many suspicious articles are caught and flagged immediately at the source. Note that flagging is not restricted to patrol but may hap-

[1] Occasionally users create articles about existing unimportant entities and present them as important, as in the case of a Scottish worker who created an article about himself claiming he was a highly decorated army officer [37]. We treat these cases the same way as fully fabricated ones: whether Captain Sir Alan Mcilwraith never existed or exists but is in fact a Glasgow call center employee does not make a real difference for all intents and purposes.

[2] `{{db-hoax}}` for blatant, and `{{hoax}}` for less obvious, hoaxes.

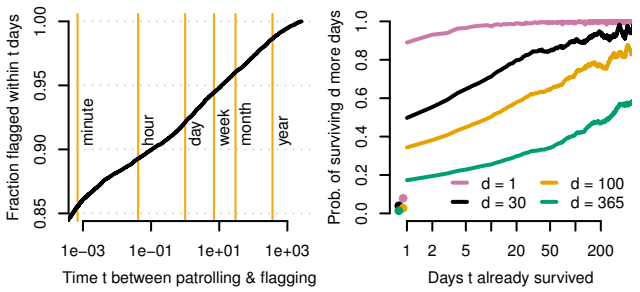


Figure 2: (a) Cumulative distribution function (CDF) of hoax survival time. Most hoaxes are caught very quickly. (b) Time the hoax has already survived on *x*-axis; probability of surviving *d* more days on *y*-axis (one curve per value of *d*). Dots in bottom left corner are prior probabilities of surviving for *d* days.

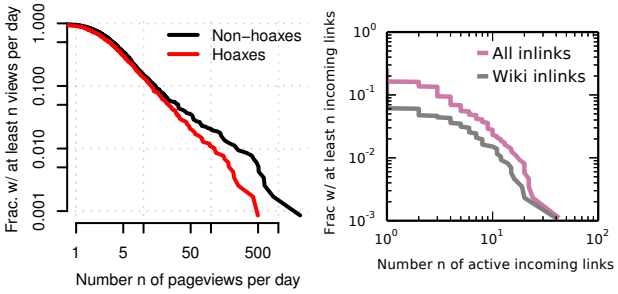


Figure 3: CCDFs of (a) number of pageviews for hoaxes and non-hoaxes (14% of hoaxes get over 10 pageviews per day during their lifetime) and (b) number of active inlinks from Web.

pen at any point during the lifetime of the article. Once flagged, the article is discussed among Wikipedians and, depending on the verdict, deleted or reinstituted (by removing the hoax template). The discussion period is generally brief: 88% of articles that are eventually deleted are deleted within a day of flagging, 95% within a week, and 99% within a month. We define the *survival time* of a hoax as the time between patrolling and flagging (Fig. 1).

In this paper we consider as hoaxes all articles of the English Wikipedia that have gone through this life cycle of creation, patrol, flagging, and deletion. There are 21,218 such articles.

3. REAL-WORLD IMPACT OF HOAXES

Disinformation is detrimental if it affects many people. The more exposure hoaxes get, the more we should care about finding and removing them. Hence, inspired by the aforementioned Jimmy Wales quote that "[t]he worst hoaxes are those which (a) last for a long time, (b) receive significant traffic, and (c) are relied upon by credible news media" [33], we quantify the impact of hoaxes with respect to how long they survive (Sec. 3.1), how often they are viewed (Sec. 3.2), and how heavily they are cited on the Web (Sec. 3.3).

3.1 Time till discovery

As mentioned in Sec. 2, since November 2007 all newly created articles have been patrolled by trusted editors. Indeed, as shown by Fig. 2(a), most of the hoaxes that are ever discovered are flagged immediately at the source: *e.g.*, 90% are flagged within one hour of (so basically, during) patrol. Thereafter, however, the detection rate slows down considerably (note the logarithmic *x*-axis of Fig. 2(a)): it takes a day to catch 92% of eventually detected hoaxes, a week to catch 94%, a month to catch 96%, and one in a hundred survives for more than a year.

Next we ask how the chance of survival changes with time. For this purpose, Fig. 2(b) plots the probability of surviving for at least $t+d$ days, given that the hoax has already survived for t days, for $d = 1, 30, 100, 365$. Although the chance of surviving the first day is very low at only 8% (Fig. 2(a)), once a hoax has survived that day, it has a 90% chance of surviving for at least another day, a 50% chance of surviving for at least one more month, and an 18% chance of surviving for at least one more year (up from a prior probability of only 1% of surviving for at least a year). After this, the survival probabilities keep increasing; the longer the hoax has already survived, the more likely it becomes to stay alive.

In summary, most hoaxes are very short-lived, but those that survive patrol have good odds of staying in Wikipedia for much longer. There is a relatively small number of longevous hoaxes, but as we show later, these hoaxes attract significant attention and a large number of pageviews.

3.2 Pageviews

Next we aim to assess the impact of Wikipedia hoaxes by studying pageview statistics as recorded in a dataset published by the Wikimedia Foundation and containing, for every hour since December 2007, how often each Wikipedia page was loaded during that hour [36].

We aggregate pageview counts for all hoaxes by day and normalize by the number of days the hoax survived, thus obtaining the average number of pageviews received per day between patrolling and flagging. Since this quantity may be noisy for very short survival times, we consider only hoaxes that survived for at least 7 days.[3] This leaves us with 1,175 of the original 21,218 hoaxes.

The complementary cumulative distribution function (CCDF) of the average number of pageviews per day is displayed as a red line in Fig. 3(a). As expected, we are dealing with a heavy-tailed distribution: most hoaxes are rarely viewed (median 3 views per day; 86% get fewer than 10 views per day), but a non-negligible number get a lot of views; *e.g.*, 1% of hoaxes surviving for at least a week get 100 or more views per day on average. Overall, hoaxes are viewed less than non-hoaxes, as shown by the black line in Fig. 3(a) (median 3.5 views per day; 85% get fewer than 10 views per day; for each hoax, we sampled one random non-hoax created on the same day as the hoax).

The facts that (1) some hoaxes survive much longer than others (Fig. 2(a)) and (2) some are viewed much more frequently per day than others (Fig. 3(a)) warrant the hypothesis that hoaxes might have a constant expected total number of pageviews until they are caught. This hypothesis would predict that plotting the total lifetime number of pageviews received by hoaxes against their survival times would result in a flat line. Fig. 4(a) shows that this is not the case, but that, instead, the hoaxes that survive longer also receive more pageviews.[4]

[3]To avoid counting pageviews stemming from patrolling and flagging, we start counting days 24 hours after the end of the day of patrolling, and stop counting 24 hours before the start of the day of flagging.

[4]It could be objected that this might be due to a constant amount of bot traffic per day (which is not excluded from the pageview dataset we use). To rule this out, we assumed a constant number b of bot hits per day, subtracted it (multiplied with the survival time) from each hoax's total count, and repeated Fig. 4(a) (for various values of b). We still observed the same trend (not plotted for space reasons), so we conclude that Fig. 4(a) is not an artifact of bot traffic.

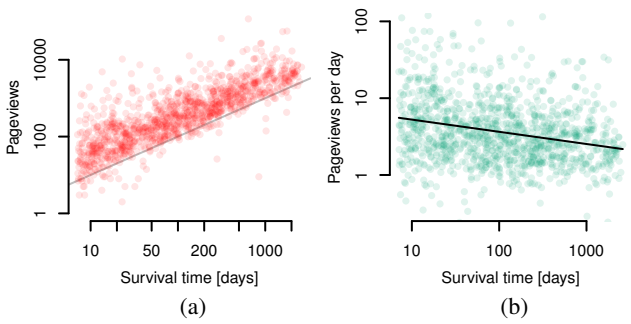


Figure 4: Longevous hoaxes are (a) viewed more over their lifetime (gray line $y = x$ plotted for orientation; not a fit) and (b) viewed less frequently per day on average (black line: linear-regression fit).

Finally, when plotting survival times against per-day (rather than total) pageview counts (Fig. 4(b)), we observe a negative trend (Spearman correlation -0.23). That is, pages that survive for very long receive fewer pageviews per day (and *vice versa*).

Together we conclude that, while there is a slight trend that hoaxes with more daily traffic generally get caught faster (Fig. 4(b)), it is not true that hoaxes are caught after a constant expected number of pageviews (Fig. 4(a)). It is not the case that only obscure, practically never visited hoaxes survive the longest; instead, we find that some carefully crafted hoaxes stay in Wikipedia for months or even years and get over 10,000 pageviews (24 hoaxes had over 10,000 views, and 375 had over 1,000 views).

3.3 References from the Web

Next we aim to investigate how different pages on the Web link and drive traffic to the hoax articles. While in principle there may be many pages on the Web linking to a particular Wikipedia hoax, we focus our attention on those links that are actually traversed and bring people to the hoax. To this end we utilize 5 months' worth of Wikipedia web server logs and rely on the HTTP referral information to identify sources of links that point to Wikipedia hoaxes.

In our analysis we only consider the traffic received by the hoax during the time it was live on Wikipedia, and not pre-creation or post-deletion traffic. There are 862 hoax articles that could potentially have received traffic during the time spanned by the server logs we use. We filter the logs to remove traffic that may have been due to article creation, patrol, flagging, and deletion, by removing all those requests made to the article during a one-day period around these events. This gives us 213 articles, viewed 23,353 times in total. Furthermore, we also categorize the different sources of requests into five broad categories based on the referrer URL: search engines, Wikipedia, social networks (Facebook and Twitter), Reddit, and a generic category containing all others. We define all search engine requests for an article as representing a single inlink. For the other categories, the inlink is defined by the URL's domain and path portions. We show the CCDF of the number of inlinks for the hoax articles in Fig. 3(b). On average, each hoax article has 1.1 inlinks. Not surprisingly, this distribution is heavily skewed, with most articles having no inlinks (median 0; 84% having at most one inlink). However, there is a significant fraction of articles with more inlinks; *e.g.*, 7% have 5 or more inlinks.

Table 1 gives the distribution of inlinks from different sources. Among the articles that have at least one inlink, search engines, Wikipedia, and "others" are the major sources of inbound connections, providing 35%, 29%, and 33% of article inlinks on average.

Metric	SE	Wiki	SN	Reddit	Others
Average inlinks	0.78	2.1	0.08	0.15	1.3
Median inlinks	1	1	0	0	1
Inlinks per article	35%	29%	0.6%	3%	33%

Table 1: Number of inlinks per hoax article ("SE" stands for search engines, "SN" for social networks).

These hoax articles have 2.1 inlinks from Wikipedia and 1.3 from "other" sources on average.

Overall, the analysis indicates that the hoax articles are accessible from multiple different locations, increasing the chances that they are viewed. Moreover, hoaxes are also frequently reached through search engines, indicating easy accessibility.

4. CHARACTERISTICS OF SUCCESSFUL HOAXES

In the present section we attempt to elicit typical characteristics of Wikipedia hoaxes. In particular, we aim to gain a better understanding of (1) how hoaxes differ from legitimate articles, (2) how successful hoaxes differ from failed hoaxes, and (3) what features make a legitimate article be mistaken for a hoax.

To this end we compare four groups of Wikipedia articles in a descriptive analysis:

1. *Successful hoaxes* passed patrol, survived for significant time (at least one month from creation to flagging), and were frequently viewed (at least 5 times per day on average).
2. *Failed hoaxes* were flagged and deleted during patrol.
3. *Wrongly flagged articles* were temporarily flagged as hoaxes, but were acquitted during the discussion period and were hence not deleted.
4. *Legitimate articles* were never flagged as hoaxes.

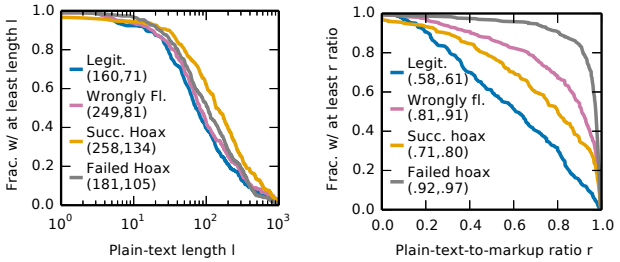
The set of all successful hoaxes consists of 301 pages created over a period of over 7 years. The usage patterns and community norms of Wikipedia may have changed during that period, and we want to make sure to not be affected by such temporal variation. Hence we ensure that the distribution of creation times is identical across all four article groups by subsampling an equal number of articles from each of groups 2, 3, and 4 while ensuring that for each successful hoax from group 1 there is another article in each group that was created on the same day as the hoax.

Given this dataset, we investigate commonalities and differences between the four article groups with respect to four types of features: (1) Appearance features (Sec. 4.1) are properties of the article that are immediately visible to a reader of the article. (2) Network features (Sec. 4.2) are derived from the so-called ego network formed by the other articles linked from the article in question. (3) Support features (Sec. 4.3) pertain to mentions of the considered article's title in other articles. (4) Editor features (Sec. 4.4) are obtained from the editor's activity before creating the article in question.

4.1 Appearance features

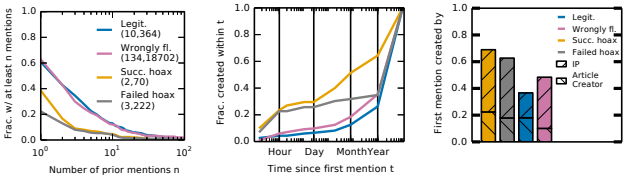
We use the term *appearance features* to refer to characteristics of an article that are directly visible to a reader.

Plain-text length. One of the most obvious features of an article is its length, which we define as the number of content words, obtained by first removing wiki markup (templates, images, references, links to other articles and to external URLs, etc.) from the article source text and then tokenizing the resulting plain text at word boundaries.



(a) Plain-text length (b) Plain-text-to-markup ratio

Figure 5: CCDFs of appearance features; means and medians in brackets.



(a) Prior mentions (b) 1st prior mention (c) 1st-men. creator

Figure 7: Support features: (a) CCDF of number of mentions prior to article creation (means/medians in brackets). (b) CDF of time from first prior mention to article creation. (c) Probability of first prior mention being inserted by hoax creator or anonymous user (identified by IP address), respectively.

Fig. 5(a) demonstrates that successful hoaxes are particularly long: their median number of content words is 134 and thus nearly twice as large as the median of legitimate articles (71). Further, and maybe surprisingly, failed hoaxes are the second most verbose group: with a median of 105 words, they are nearly 50% longer than legitimate articles.

Plain-text-to-markup ratio. While the length of an article is largely determined by the plain-text words it contains, the overall visual appearance of an article depends heavily on the wiki markup contained in the article source. The ratio of words after *vs.* before markup stripping may be taken as a measure of how "wiki-like" the article is: a ratio of 1 implies no wiki markup, *i.e.*, no wiki links to other articles, no infoboxes, references, images, footnotes, etc., and the article would look rather different from a typical Wikipedia article; the smaller the ratio, the more effort was devoted to making the article conform to Wikipedia's editorial standards.

Fig. 5(b) reveals striking differences between article groups. On one extreme, legitimate articles contain on average 58% plain text. On the other extreme, failed hoaxes consist nearly entirely of plain text (92% in the mean). Successful hoaxes and wrongly flagged articles take a middle ground.

This suggests that embellishing a hoax with markup increases its chances of passing for legitimate and that, conversely, even legitimate articles that do not adhere to the typical Wikipedia style are likely to be mistaken for hoaxes. It is not so much the amount of bare-bones content that matters—wrongly flagged articles (median 81 words; Fig. 5(a)) are similarly long to unflagged legitimate articles (median 71)—but rather the amount of mixed-in markup.

Wiki-link density. Links between articles (so-called *wiki links*) constitute a particularly important type of markup. Legitimate articles and successful hoaxes contain similar numbers of wiki links (Fig. 6(a); medians 12 and 11, respectively); failed hoaxes and wrongly flagged articles, on the other hand, are much less well connected: their medians are only 0 and 3, respectively.

While the number of wiki links is similar for legitimate articles and hoaxes, we saw previously that successful hoaxes are nearly twice as long as legitimate articles on average. Hence another interesting measure is the density of wiki links, defined here as the number of wiki links per 100 words (counted before markup stripping because wiki links may be embedded into markup such as templates).

Under this measure the picture changes: as evident in Fig. 6(b), successful hoaxes have significantly fewer outlinks per 100 words than legitimate articles (medians 5 *vs.* 7). Wrongly flagged articles (median 2) look again more like hoaxes than legitimate articles, which is probably a contributing factor to their being suspected to be hoaxes.

Web-link density. Considering the density of links to general Web resources (Fig. 6(c)), rather than to other Wikipedia articles, results in the same conclusion that legitimate articles are considerably better referenced than articles that are, or are mistaken for, hoaxes.

4.2 Link network features

Above we treated features derived from embedded links as appearance features, since the links are clearly visible to a reader of the article. But they are at the same time features of the hyperlink network underlying Wikipedia. While outlinks constitute a first-order network feature (in the sense that they deal only with direct connections to other articles), it is also interesting to consider higher-order network features, by looking not only at what the article is connected to, but also how those connected articles are linked amongst each other.

Ego-network clustering coefficient. To formalize this notion, we consider an article A's *ego network,* defined as the undirected graph spanned by the articles linked from A and the links between them (A itself is not included in the ego network). Given the ego network, we compute its clustering coefficient [35] as the actual number of edges in the ego network, divided by the number of edges it would contain if it were fully connected.

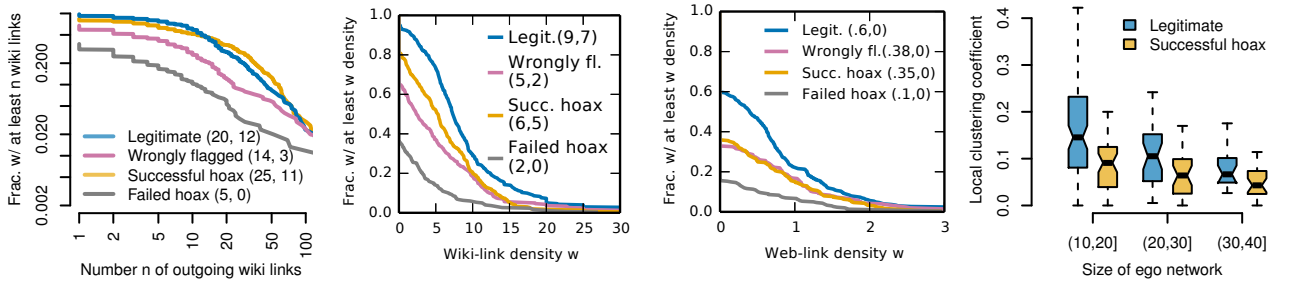
Fig. 6(d) shows that legitimate articles tend to have larger clustering coefficients than successful hoaxes, which implies that their outlinks are more coherent. It appears to be difficult to craft a fake concept that is embedded into the network of true concepts in a realistic way. In other words, making an article look realistic on the surface is easy; creating a realistic network fingerprint is hard.

As an aside, Fig. 6(d) is stratified by ego-network size because otherwise clustering coefficient and ego-network size could be confounded, as shown by the negative trend: when an article links to many other articles, they tend to be less tightly connected than when it links to only a few selected other articles—akin to a precision/recall tradeoff.

4.3 Support features

Something completely fabricated should never have been referred to before it was invented. Therefore we expect the frequency with which an article's name appears in other Wikipedia articles before it is created to be a good indicator of whether the article is a hoax.

Number of prior mentions. To test this hypothesis, we process Wikipedia's entire revision history (11 terabytes of uncompressed text) and, for each article A included in one of our four groups, identify all revisions from before A's creation time that contain A's title as a substring.



(a) Number of wiki links (b) Wiki-link density (c) Web-link density (d) Ego-net clustering coefficient

Figure 6: Link characteristics: CCDFs (means/medians in brackets) of (a) number of wiki links, (b) wiki-link density, and (c) Web-link density. (d) Ego-network clustering coefficient as function of ego-network size (nodes of outdegree at most 10 neglected because clustering coefficient is too noisy for very small ego networks; nodes of outdegree above 40 neglected because they are very rare).

Of course, such a crude detector is bound to produce false positives.[5] But since the false-positive rate is likely to be similar across groups of articles, it is nonetheless useful for comparing different groups in a relative fashion, as done in Fig. 7(a), which shows that the two types of non-hoaxes (wrongly flagged and unflagged, *i.e.*, legitimate) have very similar distributions of prior mentions; analogously, the two types of hoaxes (successful and failed) resemble each other. One important difference between successful and failed hoaxes, however, is that of the successful ones, 40% are mentioned in at least one other article before creation, whereas this is the case for only 20% of the failed ones. (At 60% the rate is much higher for non-hoaxes.)

Time of first prior mention. Part of the reason why so many hoaxes have a mention before creation is due to the aforementioned false-positive rate of our simplistic mention detector. But there is a second reason: smart hoaxsters may carefully prepare the environment for the launch of their fabrication by planting spurious mentions in other articles, which creates an illusion of external support.

Consider Fig. 7(b), which plots the cumulative distribution function of the time between the appearance of the first mention of an article A in some other article and the creation of A itself. Legitimate articles are referred to long before they are created: 75% have been mentioned for over a year by the time the article is created, and under 5% have been mentioned for less than an hour. Successful hoaxes, on the contrary, have a probability of only 35% of having been mentioned for over a year when the hoax is created,[6] and a probability of 24% of having been mentioned for less than an hour—up by a factor of about 5 compared to non-hoaxes. We suspect that it is in many cases the hoaxster herself who inserts the first mention so briefly before creating the hoax in order to lend it artificial support.

Creator of first prior mention. Looking for additional evidence for this hypothesis, we explicitly investigate who is responsible for the first mention. To this end, Fig. 7(c) plots the fraction of first mentions made by the article creator herself. (Recall from Sec. 2 that we always know which user created an article, since anonymous users do not have permission to create new articles.) We expected most hoaxes to have been first mentioned by the hoaxster herself, but inspecting the figure we see that this is not the case: the fraction of first mentions inserted by the article creator is only slightly larger for hoaxes than for non-hoaxes (21% *vs.* 19%).

It seems that hoaxsters are smarter than that: Fig. 7(c) also tells us that 45% of first mentions are introduced by non-logged-in users identified only by their IP address, whereas the baseline over legitimate articles is only 19% here. Hence it seems likely that the anonymous user adding the first mention is often the hoaxster herself acting *incognito,* in order to leave no suspicious traces behind.

We conjecture that a significant fraction of first mentions from logged-in users other than the hoaxsters in fact stem from the hoaxsters, too, via fake "sockpuppet" accounts, but we have no means of verifying this hypothesis.

4.4 Editor features

The evidence from the last subsection that hoaxsters may act undercover to lend support to their fabricated articles motivates us to take a broader look at the edit histories of article creators.

Number of prior edits and editor age. Consider Fig. 8, where we explore article creators' edit histories under two metrics: the time gone by since the user registered on Wikipedia (Fig. 8(a)) and the number of edits they have made prior to creating the article in question (Fig. 8(b)).

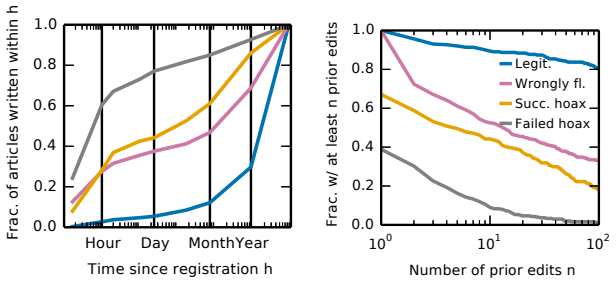
The originators of typical legitimate articles are established members of the Wikipedia community: three-quarters of all such articles were started by editors who have been registered for more than a year, with a median of over 500 prior edits.[7] On the contrary, the three groups of articles that are flagged as hoaxes (whether they really are hoaxes or not) are created by much more recent accounts, in the following order: failed-hoax authors are the youngest members, followed by the creators of successful hoaxes, and finally by those of articles flagged wrongly as hoaxes.

In particular, while only about 3% of legitimate-article authors create the article within the hour of registration, the fractions are 60% for creators of failed hoaxes, and 25% for those of successful hoaxes and wrongly flagged articles. In the case of wrongly flagged articles, we suspect that inexperience may cause users to write articles that do not comply with Wikipedia's standards (*cf.* Fig. 5). This in combination with the concern that, due to the recent registration date, the account might have been created specifically for creating the hoax might lead patrollers to erroneously suspect the new article of having been fabricated.

[5]For instance, a mention of the newspaper *The Times* will be spuriously detected in the Bob Dylan article because it mentions the song *The Times They Are a-Changin*.

[6]This number is much larger for failed hoaxes, which begs an explanation. Eyeballing the data, we conjecture that this is caused by obvious, failed hoaxes often being created with mundane and commonplace names, such as "French immigrants" or "Texas style".

[7]In order to limit the number of calls to the Wikipedia API, we collected at most 500 edits per user. Therefore, the median measured in this setting (500) is a lower bound of the real median.



(a) Time since registration (b) Number of prior edits

Figure 8: Editor features: (a) CDF of time between account registration and article creation. (b) CCDF of number of edits by same user before article creation.

Feature	Group
Plain-text length	Appearance (Sec. 4.1)
Plain-text-to-markup ratio	Appearance
Wiki-link density	Appearance
Web-link density	Appearance
Ego-network clustering coefficient	Network (Sec. 4.2)
Number of prior mentions	Support (Sec. 4.3)
Time of first prior mention	Support
Creator of first prior mention	Support
Number of prior edits	Editor (Sec. 4.4)
Editor age	Editor

Table 2: Features used in the random-forest classifiers.

5. AUTOMATIC HOAX DETECTION

Having gained several valuable insights on the characteristics of Wikipedia hoaxes and their differences from other types of articles, we are now in a position to apply these findings by building machine-learned classifiers to automate some important decisions revolving around hoaxes. We consider the following four tasks:

1. Will a hoax get past patrol?
2. How long will a hoax survive?
3. Is an article a hoax?
4. Is an article flagged as such really a hoax?

The first two tasks take the hoaxster's perspective and ask how high the chances are of the hoax being successful. The latter two tasks take the patrollers' perspective and aim to help them make an accurate decision during patrol and after.

All classifiers use the same algorithm and features, but are fitted on different training sets of positive and negative examples. This allows us to analyze the fitted weights in order to understand what features matter most in each task.

Classification algorithm. We experimented with a variety of classification algorithms—logistic regression, support vector machines, and random forests—and found the latter to work best. Hence all results reported here were obtained using random forests [4].

We use balanced training and test sets containing equal numbers of positive and negative examples, so random guessing results in an accuracy, as well as an area under the receiver operating characteristic (ROC) curve (AUC) of 50%.

Features. All features used by the classifier have been discussed in detail in Sec. 4 and are summarized in Table 2.

In the rest of this section we provide more details on each of the four tasks (Sec. 5.1) and then move on to presenting and discussing the results we obtained (Sec. 5.2).

5.1 Classification tasks

Task 1: Will a hoax get past patrol? Here the objective is to predict if a hoax will pass the first hurdle in its life cycle (Fig. 1), *i.e.*, if it will manage to trick the patroller into believing that it is a legitimate article.

Such a classifier could tell the hoaxster whether the hoax is ready to be submitted to the patrolling process yet. It would also be useful from the patroller's perspective because the fitted feature weights can give us insights into which features make a hoax slip through patrol; we could then counteract by scrutinizing those characteristics more carefully.

Here the set of positive examples consists of all 2,692 hoaxes that were not flagged by the users who patrolled them. The negative examples are sampled randomly from the set of 12,901 hoaxes that are correctly flagged by the patroller, while ensuring that for each positive article we have a negative article created on the same day.

Task 2: How long will a hoax survive? Our second task is to predict the survival time of hoaxes that have managed to pass patrol, defined as the time between patrol and flagging (Fig. 1). We phrase this as a binary decision problem by fixing a threshold τ and asking whether a hoax will survive for at least τ minutes. We repeat this task for various values of τ, ranging from one minute to one year.

Given τ, the positive examples are all hoaxes that survived for at least τ minutes from patrol to flagging. The negative set consists of hoaxes flagged within τ minutes from patrol. The larger of the two sets for the given τ is subsampled to match the smaller set in size.

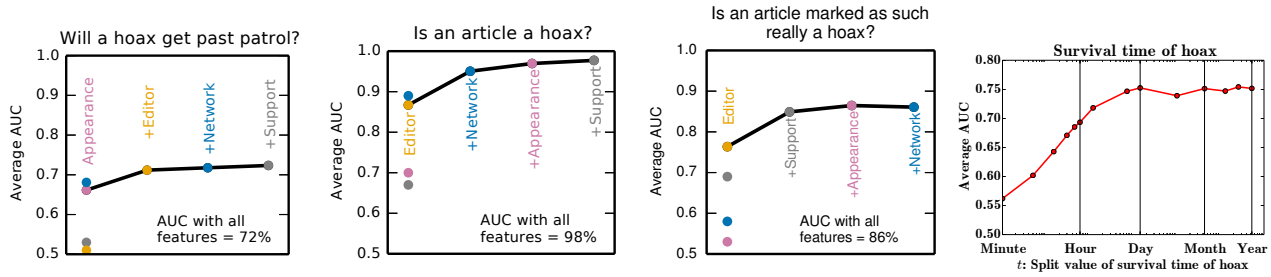
Task 3: Is an article a hoax? In this task, the classifier is supposed to assess if an article that has passed patrol is a hoax or not. In the language of Fig. 1, the task aims to automate the flagging step. This classifier could be employed to double-check the decisions made by human patrollers and thereby decrease their false-negative rate.

Here the positive examples are the 2,692 articles that passed patrol but were later flagged as hoaxes and deleted. As argued in the introduction, the most detrimental hoaxes are those that survive for a long time and attract significant traffic. In order to equip our classifier with the ability to detect this subclass, we include only those 301 hoaxes as positive examples that have existed for at least 30 days from creation to flagging and that have received an average of at least 5 pageviews during this time. For each hoax in the positive set we randomly sample one negative example from among all articles that were created on the same day as the hoax and were never flagged or deleted.

Task 4: Is an article marked as such really a hoax? The final classification task deals with the scenario in which an article has been flagged as a hoax by a Wikipedia user, and our goal is to double-check if the article is indeed a hoax. That is, this classifier is supposed to act as a safeguard between the flagging and deletion steps (Fig. 1).

In other words, while task 3 aims to decrease human patrollers' false-negative rate, the classifier developed here may decrease their false-positive rate. This could be very valuable because false positives come at a large cost: if an article is unjustly deleted as a hoax, this might discourage the editor to contribute further to Wikipedia.

The negative set comprises the 960 articles that were wrongly flagged, *i.e.*, that were later acquitted by having the hoax flag removed and were never deleted. Candidates for positive examples are all articles that were flagged as hoaxes and eventually deleted. To create a balanced dataset, we pair each negative example with a positive examples whose creation and flagging dates are closely aligned with those of the negative example (we use propensity score matching [31] to perform the pairing).



(a) Task 1 (b) Task 3 (c) Task 4 (d) Task 2

Figure 9: (a–c) Results of forward feature selection for tasks 1, 3, 4. (d) Performance (AUC) on task 2 as function of threshold τ.

5.2 Results

Table 3 reports the performance on tasks 1, 3, and 4 when using all features of Table 2. Task 2 depends on the threshold τ, so we plot the AUC as function of τ in Fig. 9(d).

	Task	Acc.	AUC
1	Will a hoax get past patrol?	66%	71%
3	Is an article a hoax?	92%	98%
4	Is an article flagged as such really a hoax?	76%	86%

Table 3: Classification results; for task 2, *cf.* Fig. 9(d).

Maybe surprisingly, deciding if an article is a hoax or not (task 3) is the easiest task, with an accuracy (AUC) of 92% (98%). Performance is also quite high on the task of deciding whether something that has been flagged as a hoax is really one (task 4); here we achieve an accuracy (AUC) of 76% (86%). The hardest tasks are to predict if a hoax will pass patrol (task 1; accuracy 66%, AUC 71%) and how long it will survive once it has passed patrol (task 2): Fig. 9(d) shows that the AUC increases with the threshold τ, but levels off at 75% around $\tau = 1$ day. That is, one day seems to be a natural threshold that separates successful from failed hoaxes. This echoes our finding from Fig. 2(b), where we saw that surviving the first day immensely boosts the odds of surviving for longer.

Feature importance. In order to understand which features are important for which task, we evaluate smaller models that consist of only one of the four feature groups (Table 2). The performance of these smaller models is shown by the vertically aligned dots in the leftmost columns of Fig. 9(a)–9(c). For tasks 3 and 4, which deal with deciding if something is a hoax, features of the creator's edit history are most effective; on task 3 (hoax *vs.* non-hoax), the network feature (ego-network clustering coefficient) does equally well. Task 1, where we predict if a given hoax will pass patrol, profits most from appearance and editor features.

Next, we perform forward feature selection to understand what the marginal values of additional features are. The results are plotted as the black curves in Fig. 9(a)–9(c).[8] The conclusion is that all feature groups contribute their share, but with diminishing returns.

Trawling Wikipedia for hoaxes. In order to find hoaxes that are still present in Wikipedia, we deployed the hoax-*vs.*-non-hoax classifier on Wikipedia's entire revision history. We discuss the results in detail online.[9] To give but two examples, our algorithm identified the article about "Steve Moertel", an alleged Cairo-born U.S. popcorn entrepreneur, as a hoax. The article was deleted by an editor who confirmed the article's hoax status after we had flagged it—and after it had survived in Wikipedia for 6 years and 11 months. Similarly, we flagged the article about "Maurice Foxell", an alleged children's book author and Knight Commander of the Royal Victorian Order; the article was deleted by an editor after it had survived for 1 year and 7 months.

[8]We performed forward feature selection on the training set and report performance on the testing set. This is why the first selected feature may have lower performance than other features.

[9]http://snap.stanford.edu/hoax/

6. HUMAN GUESSING EXPERIMENT

The observational analysis of Sec. 4 allowed us to gain many insights, but it also has some shortcomings. First, survival time defined by the period between patrol and flagging is not a perfect indicator of the quality of a hoax, as the hoax may have survived for a long time for a variety of reasons; *e.g.*, it may be the case that the false information is disguised in a truly skillful manner, or simply that it was sloppily patrolled and was afterwards seen by only few readers who could have become suspicious. So by only considering the observational data we have analyzed above, we cannot know which hoax survived for which reason.

Second, the binary label whether a hoax passed patrol or not is not necessarily representative of how likely a regular Wikipedia reader, rather than a patroller, would be to believe the hoax. Patrollers are encouraged to base their decision on all available information, including fact-checking on the Web via search engines, verifying included references, inspecting the article creator's edit history, etc. We suspect that most Wikipedia readers do not use such devices during casual reading and are therefore more likely to fall prey to a hoax that looks legitimate on the surface.

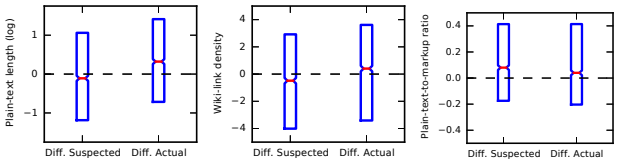
To overcome these shortcomings and understand what makes a hoax credible to average readers rather than patrollers, we now complement our observational findings with an experiment. The idea is to (1) create an identical situation of scrutiny across a variety of hoaxes, thus mitigating the first concern from above, and (2) disallow the use of external resources such as search engines, thus addressing the second concern.

6.1 Methodology

In designing the experiment, we start by selecting 64 successful hoaxes according to the definition from the beginning of Sec. 4. We then create an equally sized set of legitimate, non-hoax articles such that (1) for each hoax we have a legitimate article created on the same day as the hoax and (2) the two sets have nearly identical distributions of the appearance features of Sec. 4.1, which we achieve via propensity score matching [31].[10]

We then created 320 random hoax/non-hoax pairs such that each hoax was paired with 5 distinct non-hoaxes and *vice versa*. These

[10]We additionally balance the sets with respect to the numbers of sections, images, and references in the articles.



(a) Plain-text length (b) Wiki-link dens. (c) Plain-text/markup

Figure 10: Human bias in the guessing experiment with respect to three appearance features *f*. Left boxes: difference δ of suspected hoax minus suspected non-hoax. Right boxes: difference δ^* of actual hoax minus actual non-hoax.

pairs were then shown side-by-side in random order to human raters on Amazon Mechanical Turk, who were asked to decide which of the two articles is a hoax by only looking at the text and not searching the Web. Each pair was given to 10 raters, so we collected 3,200 labels in total (50 per hoax). We assured the quality of raters as described in the appendix.

6.2 Results

Human *vs.* classifier accuracy. Human accuracy on all rated pairs is 66%. The macro-average that gives equal weight to all users (hoaxes) is 63% (66%). Given that random guessing on the task would give 50%, this performance is surprisingly weak.[11] In comparison, we tested our hoax-*vs.*-non-hoax classifier (task 3 of Sec. 5) on the same pairs shown to humans and achieved an accuracy of 86%, thus outperforming humans by a large margin.[12]

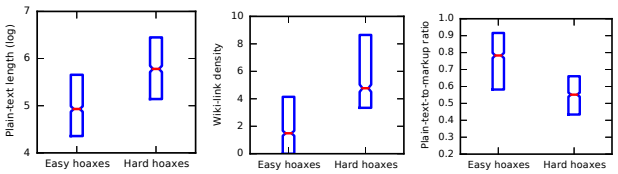
This classifier used all features of Sec. 5. The human, however, saw only the articles themselves and was not allowed (and for most features not even able to) take network, support, and editor features into account. To allow for a fairer comparison, we therefore also tested a version of our classifier that uses only appearance features, obtaining an accuracy of only 47%. This weak (roughly random) performance is to be expected, since the sets of hoaxes and non-hoaxes were constructed to have very similar distributions with respect to appearance features (*cf.* above), so these features should be uninformative for the task.

We conclude that features that look beyond the surface, such as the article creator's edit history, the mentions received from other articles, and the density of the article's ego network, are of crucial importance for deciding whether an article is a hoax: they make the difference between random and above-human performance.

Human bias. Our next goal is to understand what factors humans go by when deciding what is a hoax. We proceed as follows: given a feature f of interest (such as plain-text length), compute the within-pair difference δ of the *suspected* hoax minus the suspected non-hoax for each pair. Similarly, compute the difference δ^* of the *actual* hoax minus the actual non-hoax, and compare the distributions of δ and δ^*. Now, if δ tends to be lower than δ^*, this implies that humans tend to think that lower values of f indicate hoaxes, although they would have had to choose the higher values more frequently in order to guess perfectly; in other words, they are biased to believe that articles with lower values of f are hoaxes.

[11] One might object that humans possibly *did* guess randomly, but we guarded against this via the quality-assurance mechanism described in the appendix.

[12] Since testing is done on pairs, we also trained the classifier on pairs: as the feature vector for a pair, we use the difference of the feature vectors of the left and right articles, and the classifier is tasked to predict whether the left or right article is the hoax. The training pairs did not contain articles appearing in the test pairs.



(a) Plain-text length (b) Wiki-link dens. (c) Plain-text/markup

Figure 11: Comparison of easy- and hard-to-identify hoaxes with respect to three appearance features.

Our findings from this analysis are displayed in the boxplots of Fig. 10. Here, the left box of each subfigure summarizes the distribution of δ, and the right box, that of δ^*. For instance, Fig. 10(a) shows that the suspected hoax tends to be shorter than the suspected non-hoax, whereas the actual hoax tends to be longer than the actual non-hoax. So humans have a bias towards suspecting short articles to be hoaxes that is not warranted by the dataset at hand. Similarly, we find that humans are led to believe that articles with a lower wiki-link density (Fig. 10(b)) and, to a lesser extent, with a higher plain-text-to-markup ratio (*i.e.*, less wiki markup; Fig. 10(c)), are hoaxes. Flipped around, from the hoaxster's perspective this means that a hoax stands a higher chance of succeeding if it is longer and looks more like a typical Wikipedia article.

Next we create two groups of hoaxes: those that are easy, and those that are hard, to detect for humans. To define these groups we first rank all hoaxes in increasing order according to the probability with which humans identified them correctly; the upper third then defines the easy, and the lower third the hard, cases. For each feature we then compare the distributions within the two groups. The results, shown in Fig. 11, indicate that the log median number of plain-text words of the hard group is higher by about 1 than that for the easy group, *i.e.*, hard-to-recognize hoaxes are in the (non-log) median about $e^1 \approx 2.7$ times as long as easy-to-recognize hoaxes. Similarly, hoaxes with many wiki links (Fig. 11(b)) and a low plain-text-to-markup ratio (Fig. 10(c)), *i.e.*, with many wiki-specific elements, are difficult to recognize.

Examples. Of course, it is not only simple structural and superficial features such as the length, link density, and presence of wiki-specific elements that determine if an article is recognized as a hoax. It is also, and to a large extent, the semantic content of the information conveyed that matters. Therefore we conclude our discussion of the human experiment with some qualitative remarks. Table 4 lists the hardest (top) and easiest (bottom) hoaxes (left) and non-hoaxes (right) for humans to identify correctly, where "hardness" is captured by the fraction of humans who failed to identify the article correctly across all pairs it appeared in. Hard-to-identify hoaxes are often elaborate articles about fake people, whereas the easy ones are oftentimes already given away by their titles.

The non-hoaxes that were least credible to raters frequently have titles that sound tongue-in-cheek. The article on the (real) Philippine radio station DXMM might have been mistaken so often because the version used in the experiment was very short and had no wiki links and sections, or because it was clumsily phrased, calling the station "the fruit of missions made by the Missionary Oblates of Mary Immaculate in the difficult and harsh fields of Mindanao and Sulu archipelago in southern Philippines."

7. RELATED WORK

Hoaxes on Wikipedia are an example of *disinformation* [17, 11]. Wikipedia defines disinformation as "intentionally false or inaccu-

Acc.	Hoax
0.333	TV5 (Malaysia)
0.341	Tom Prescillo
0.362	Alexander Ivanovich Popov
0.391	Noah Chazzman
0.400	Dav Sorado
...	...
0.867	The Oregon Song
0.875	Nicktoons: Dark Snap
0.884	Breast Touching Festival of China
0.955	Burger King Stunners
0.957	Mama Mo Yeah

Acc.	Non-hoax
0.292	Țițeica
0.312	DXMM
0.364	Better Made Potato Chips Inc.
0.370	Olympiacos B.C. vs Punch Delft (prehistory)
0.378	Don't Come Home for Christmas
...	...
0.872	List of governors of Islamic Egypt
0.891	Bobby Brown discography
0.907	List of Naruto episodes (season 4)
0.957	Alpine skiing at the 2002 Winter Olympics – Women's slalom
0.958	USS Charles P. Crawford (SP-366)

Table 4: Hoaxes (left) and non-hoaxes (right) that were hardest (top) and easiest (bottom) for humans to identify correctly.

rate information that is spread deliberately. It is an act of deception and false statements to convince someone of untruth." Disinformation is frequently distinguished from *misinformation,* which is information that is unintentionally false.

Several pieces of related work analyze the impact of false information on Web users. In particular, a number of papers [25, 34, 19] investigate which factors boost or hurt credibility, and by which strategies users can evaluate the credibility of online sources. Such survey-based studies have been carried out both on the Web in general [12, 13, 26] as well as on Twitter in particular [28]. Our work focuses on hoaxes as an example of disinformation and adds to this line of work by showing that people do not perform particularly well when trying to distinguish false information from the truth.

When false information in the form of rumors, urban legends, and conspiracy theories appears in a social network, users are often led to share and disseminate it [7, 9]. There is a rich line of empirical investigations and case-based studies of how this propagation happens, *e.g.*, in Facebook [9, 15], Twitter [16], and Sina Weibo [42]. Additionally, researchers have proposed theoretical models of how rumors and misinformation propagate in social networks and how their spread may be contained [32, 1]. Other work has developed approximation algorithms for the problem of limiting the spread of misinformation by selecting a small number of nodes to counteract the effect of misinformation [5, 29]. Our work relates to this line of research by studying misinformation on Wikipedia and assessing its impact on the community and the broader ecosystem of the Web.

More related to our present work is prior research that aims to build automatic methods for assessing the credibility of a given set of social media posts [6, 22, 42, 43]. Most of the work in this area has focused on engineering features that allow for detecting rumorous, fake, and deceptive content in social media [22]. For example, Kwon *et al.* [21] identify temporal, structural, and linguistic features of rumors on Twitter; Gupta *et al.* [16] use social reputation and influence patterns to predict whether images being transmitted on Twitter are real or fake; and Qazvinian *et al.* [30] attempt to predict if tweets are factual or not, while also identifying sources of misinformation. There are two main differences with respect to our work: first, by working with collaboratively authored Wikipedia content, we investigate a rather different domain; and second, Wikipedia hoaxes do not spread like social media posts, but are subject to more subtle processes, involving volunteers who constantly patrol Wikipedia in order to detect and block such content.

A final line of related work aims at developing metrics and tools for assessing the quality of Wikipedia articles. Such metrics are often based on textual properties of the article such as word counts [3], or on the edit history of the article [8, 41]; most approaches, however, focus on reputation mechanisms and the interactions between articles and their contributors [23, 18, 44]. Common to all these approaches is that editor reputation has good predictive value of article quality: edits performed by low-reputation authors have a larger probability of being of poor quality [2]. It is important to note that these projects develop metrics to assess the quality of any Wikipedia article and assume that such articles are legitimate and true, while possibly not entirely complete. The work we present here, on the contrary, investigates the distinct problem of differentiating between truthful and false information on Wikipedia.

8. CONCLUSION

In this paper we investigate impact, characteristics, and detection of hoax articles on Wikipedia. We utilize a rich labeled dataset of previously discovered hoaxes and use it to assess the real-world impact of hoax articles by measuring how long they survive before being debunked, how many pageviews they receive, and how heavily they are referred to by documents on the Web. We find that the Wikipedia community is efficient at identifying hoax articles, but that there is also a small number of carefully crafted hoaxes that survive for a long time and are well cited across the Web.

We also characterize successful hoaxes by comparing them with legitimate articles and with failed hoaxes that were discovered shortly after being created. We uncover characteristic differences in terms of article structure and content, embeddedness into the rest of Wikipedia, and features of the editor who created the hoax.

We rely on these lessons to build an automatic classification system to determine whether a given article is a hoax. By combining features derived from the article's appearance, its mentions in other articles, and its creator, as well as the Wikipedia hyperlink network, our approach achieves an AUC/ROC of 98%. We also compare our automatic hoax detection tool with the performance of human evaluators and find that humans without any specialized tools are not skilled at discerning hoaxes from non-hoaxes (63% accuracy). Our experiments show that, while humans have the tendency to rely on article appearance features, those alone are not sufficient to make accurate judgments. In contrast, our algorithms are able to utilize additional signals, such as the embeddedness of the article into the rest of Wikipedia, as well as properties of the article creator, in order to accurately identify hoaxes. To turn our insights into actions, we apply our learned model to Wikipedia's entire revision history and find hoaxes that have been hidden in it for a long time.

There are many avenues for future work. Perhaps surprisingly, our experiments have shown that even by using only superficial "content" features (*e.g.*, article length, number of links) automatic methods can quite accurately identify hoaxes. Nonetheless, a more in-depth semantic analysis of hoax content would be an intriguing avenue of future research. We observe that many well-crafted hoaxes attempt to reinforce their credibility by including links to external Web resources, some of them serious, others fictional. Understanding these mechanisms of generating spurious support could further strengthen our hoax detection methods, as would a more thorough understanding of the role of sockpuppet accounts. Finally, it would be intriguing to attempt to better understand the intentions of users who create hoaxes: is their motivation the sheer joy of vandalism or the desire to make a profit of some kind? Answering such questions will help us design future information systems that are more effectively safeguarded against the creation and propagation of disinformation.

Ackowledgments. Supported in part by NSF CNS-1010921, IIS-1149837, ARO MURI, W911NF11103, W911NF1410358, W911NF09102, DARPA XDATA, SIMPLEX, SDSI, Boeing, Facebook, SAP, VW, Yahoo, and a Wikimedia Research Fellowship (Robert West). We thank the Wikimedia Foundation, and Leila Zia in particular, for granting data access.

9. REFERENCES

[1] D. Acemoglua, A. Ozdaglar, and A. ParandehGheibi. Spread of (mis)information in social networks. *Games and Economic Behavior*, 70(2):194–227, 2010.

[2] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *WWW*, 2007.

[3] J. E. Blumenstock. Size matters: Word count as a measure of quality on Wikipedia. In *WWW*, 2008.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *WWW*, 2011.

[6] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on Twitter. In *WWW*, 2011.

[7] X. Chen, S.-C. J. Sin, Y.-L. Theng, and C. S. Lee. Why do social media users share misinformation? In *JCDL*, 2015.

[8] G. de la Calzada and A. Dekhtyar. On measuring the quality of Wikipedia articles. In *WICOW*, 2010.

[9] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *PNAS*, 113(3):554–559, 2016.

[10] R. DeNardo. *The Captain's Propensity: The Andromeda Incident II*. Strategic Book Publishing, 2013.

[11] D. Fallis. A functional analysis of disinformation. *iConference*, 2014.

[12] B. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, et al. What makes web sites credible? A report on a large quantitative study. In *CHI*, 2001.

[13] B. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber. How do users evaluate the credibility of web sites? A study with over 2,500 participants. In *DUX*, 2003.

[14] H. Frankfurt. On bullshit. *Raritan Quarterly Review*, 6(2):81–100, 1986.

[15] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *ICWSM*, 2014.

[16] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking Sandy: Characterizing and identifying fake images on Twitter during hurricane Sandy. In *WWW Companion*, 2013.

[17] P. Hernon. Disinformation and misinformation through the Internet: Findings of an exploratory study. *Government Information Quarterly*, 12(2):133–139, 1995.

[18] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in Wikipedia: models and evaluation. In *CIKM*, 2007.

[19] H. Keshavarz. How credible is information on the Web: Reflections on misinformation and disinformation. *Infopreneurship Journal*, 1(2):1–17, 2014.

[20] N. Khomami. Woman dies after taking 'diet pills' bought over internet. Website, 2015. `http://www.theguardian.com/society/2015/apr/21/woman-dies-after-taking-diet-pills-bought-over-internet` (accessed Oct. 16, 2015).

[21] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *ICDM*, 2013.

[22] T. Lavergne, T. Urvoy, and F. Yvon. Detecting fake content with relative entropy scoring. In *PAN*, 2008.

[23] E.-P. Lim, B.-Q. Vuong, H. W. Lauw, and A. Sun. Measuring qualities of articles contributed by online communities. In *WI*, 2006.

[24] M. McCormick. *Atheism and the Case Against Christ*. Prometheus Books, 2012.

[25] M. J. Metzger. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *JASIST*, 58(13):2078–2091, 2007.

[26] D. Mocanu, L. Rossi, Q. Zhang, M. Karsai, and W. Quattrociocchi. Collective attention in the age of (mis)information. *Computers in Human Behavior*, 51:1198–1204, 2015.

[27] K. Morris. After a half-decade, massive Wikipedia hoax finally exposed. Website, 2013. `http://www.dailydot.com/news/wikipedia-bicholim-conflict-hoax-deleted` (accessed Oct. 16, 2015).

[28] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Tweeting is believing? Understanding microblog credibility perceptions. In *CSCW*, 2012.

[29] N. P. Nguyen, G. Yan, M. T. Thai, and S. Eidenbenz. Containment of misinformation spread in online social networks. In *WebSci*, 2012.

[30] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *EMNLP*, 2011.

[31] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[32] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *WWW Companion*, 2015.

[33] J. Wales. How frequent are Wikipedia hoaxes like the "Bicholim Conflict"? Website, 2013. `https://www.quora.com/How-frequent-are-Wikipedia-hoaxes-like-the-Bicholim-Conflict` (accessed Oct. 16, 2015).

[34] C. N. Wathen and J. Burkell. Believe it or not: Factors influencing credibility on the Web. *JASIST*, 2002.

[35] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[36] Wikimedia Foundation. Page view statistics for Wikimedia projects. Website, 2015. `https://dumps.wikimedia.org/other/pagecounts-raw` (accessed Oct. 16, 2015).

[37] Wikipedia. Alan Mcilwraith. Website, 2015. `https://en.wikipedia.org/w/index.php?title=Alan_Mcilwraith&oldid=682760877` (accessed Oct. 16, 2015).

[38] Wikipedia. Balboa Creole French. Website, 2015. `https://en.wikipedia.org/w/index.php?title=Wikipedia_talk:List_of_hoaxes_on_Wikipedia/Balboa_Creole_French&oldid=570091609` (accessed Oct. 16, 2015).

[39] Wikipedia. Do not create hoaxes. Website, 2015. `https://en.wikipedia.org/w/index.php?title=Wikipedia:Do_not_create_hoaxes&oldid=684241383` (accessed Oct. 16, 2015).

[40] Wikipedia. Wikipedia Seigenthaler biography incident. Website, 2015. https://en.wikipedia.org/w/index.php?title=Wikipedia_Seigenthaler_biography_incident&oldid=677556119 (accessed Oct. 16, 2015).

[41] T. Wöhner and R. Peters. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *WikiSym*, 2009.

[42] Q. Xu and H. Zhao. Using deep linguistic features for finding deceptive opinion spam. In *COLING*, 2012.

[43] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on Sina Weibo. In *MDS*, 2012.

[44] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing trust from revision history. In *PST*, 2006.

APPENDIX

A. QUALITY ASSURANCE IN HUMAN GUESSING EXPERIMENT

Hoax/non-hoax pairs were issued in batches of 3; one of the 3 pairs was a test pair for which we made sure it was obvious which article was legitimate, by choosing an article about a country as the non-hoax. Raters were told they would not be paid if they did not get the test pair right (they were not told which one it was). This was to incentivize raters to make a best effort on all 3 pairs and refrain from clicking randomly. It also allows us to discard answers from raters who answered less than a minimum fraction of test pairs correctly. 92% of the test questions were answered correctly, and we discard all answers from raters with a test-question accuracy below 75%, which leaves us with 2,942 of the original 3,200 pairs.