Title

Seeing Is Believing: Do People Fail to Identify Fake Images on the Web?

Authors

Kasra, Mona
Shen, Cuihua
O'Brien, James F

Publication Date

2016-10-01

Peer reviewed

Selected Papers of AoIR 2016:
The 17th Annual Conference of the
Association of Internet Researchers
Berlin, Germany / 5-8 October 2016

SEEING IS BELIEVING: DO PEOPLE FAIL TO IDENTIFY FAKE IMAGES ON THE WEB?

Mona Kasra, PhD
University of Virginia

Cuihua Shen, PhD
University of California Davis

James O'Brien, PhD
University of California Berkeley

Introduction

If a picture is worth a thousand words, then is a manipulated picture worth a thousand lies? As the abundance of hardware and software tools continues to dramatically decrease the cost and effort required to convincingly manipulate digital images, the risks and dangers associated with ill-intentioned individuals or groups easily routing doctored images through computer and social networks to cause emotional distress or to purposefully influence opinions, attitudes, and actions have never been more severe. Not only can bad actors infiltrate cyberspace to gain or attack information online, but they can also stream manipulated information, particularly doctored visual content, to inflict cognitive stress, exploit prior beliefs, or to shape and control individuals' decisions. Due to the scope and speed of information dissemination across social media websites, visual misinformation could act as a form of Social Cyber-Attack (Goolsby) capable of manipulating crowds, propagating hysteria, confusion, distress, panic, violence, and escalating chaotic mass behavior at a fast pace and on a large scale. Unfortunately, even when doctored images are eventually exposed as forgeries, their lingering impact on viewers' emotions, viewpoints and attitudes may lead to dangerous personal and/or sociopolitical outcomes. These images could even hinder humanitarian or disaster relief efforts by spreading misinformation and promoting hostility.

We know alarmingly little about the public's vulnerabilities to visual misinformation, how individuals make credible evaluations about image authenticity, or what social and cognitive heuristics they rely on for these evaluations. This paper describes results

Suggested Citation (APA): Kasra, Shen, and O'Brien (2016, October 5-8). *Seeing Is Believing: Do People Fail to Identify Fake Images on the Web?* Paper presented at AoIR 2016: The 17th Annual Conference of the Association of Internet Researchers. Berlin, Germany. Retrieved from http://graphics.berkeley.edu/papers/Kasra-SIB-2016-10 .

from an exploratory study focusing on how individuals react and respond to images that accompany online stories in Internet-enabled communication channels (social networking sites, blogs, email), as well as their ability to identify authentic or false visual information on the Web. Our exploratory study is an initial step toward addressing the need to better understand how and why people trust online images, and how people are influenced by manipulated images both when they are aware or unaware of the manipulation. Our eventual goal is to lay the grounds for new technologies that aid Internet users in developing a healthy skepticism toward the mediated visual hoaxes, scams, and misinformation that they receive online.

Study Methodology

We designed and administered an exploratory focus group study where participants were asked to evaluate and reason about a set of images paired with news and stories similar to those they might encounter on the Web.

Image creation: We first collected 44 doctored imagery that received significant media attention in recent years and created a chart to sort the manipulation techniques and methods applied to them. Studying these images allowed us to identify a few common manipulation objectives ranging from fabricating scenes of natural disasters to political propaganda involving negative or positive portrayals of characters. The process also enabled us to identify four classes of common manipulation techniques used to forge online images: composition, elimination, retouching, and misattribution.

We then created a set of image compositions spanning a range of topics, placed them into documents showing a variety of online stories and contexts, and presented them to our participants. The final compositions were presented to the focus study participants as mockups, showing the medium purportedly used to disseminate the images (Twitter, Instagram, Facebook, or email), and the source varying from reputable outlets such as BBC, FOX News, and CNN, to general social media users with few or many followers. Images accompanied different commentaries or stories and if applicable revealed the number of viewers, likes, shares, or retweets.

Focus Groups: We chose the focus group format, which has been used in similar studies (e.g., Ringel Morris et al.), because: 1) it allows participants to explore, clarify and mutually influence their point of view in a natural collective process, interacting with others just as they would in online environments; and 2) it is more cost-effective than individual interviews yet still gives us access to multiple perspectives.

To ensure that the results of the exploratory study were minimally biased by the political climate surrounding the participants, we conducted our focus groups sessions at two separate sites, a large public university in California and a large public university in Texas. We also asked participants to complete the initial credibility rating before starting the group discussions to minimize group polarization effects and prevent the group discussion from being excessively influenced by the most outspoken subjects.

Findings

Overall, we found that participants in this pilot study performed poorly at identifying the fake online images presented to them. The questionnaire asked to what extent participants were confident about the authenticity of each image, with 1=not at all confident, 5=extremely confident. The average rating for the eleven images was 2.7, even though each of the images had been manipulated in a way that substantially changed its content.

After completing the questionnaire, there was a discussion within each focus group. These discussions revealed some common patterns. First, we found that the participants made judgments based mostly on non-image cues. These cues included the disseminating source of the image story, the media platform used, and/or captions and commentaries accompanying the images. Image-specific cues, such as inconsistencies in lighting and shadows, were rarely mentioned.

Aside from the source, the textual description or commentaries accompanying the images also appeared to play a key role in credibility evaluation or detecting a forgery. In analyzing the images, participants tended to make an assessment based on predispositions and the textual information. After the initial gravitation toward the words and applying preconceptions in examining image authenticity, they then applied post hoc analysis to look for evidence and cues that supported their assessment. In other words, if a participant wanted to believe an image was authentic or forged, they would find ways to justify their view.

Another finding was that when participants purposefully looked for clues to dismiss the authenticity of an image due to existing predispositions, they tended to fail to identify the elements of the image compositions that had actually been modified. In many cases, the participants were unable to identify any manipulation, and their solution was to assume misattribution rather than forgery.

Discussion & Conclusion

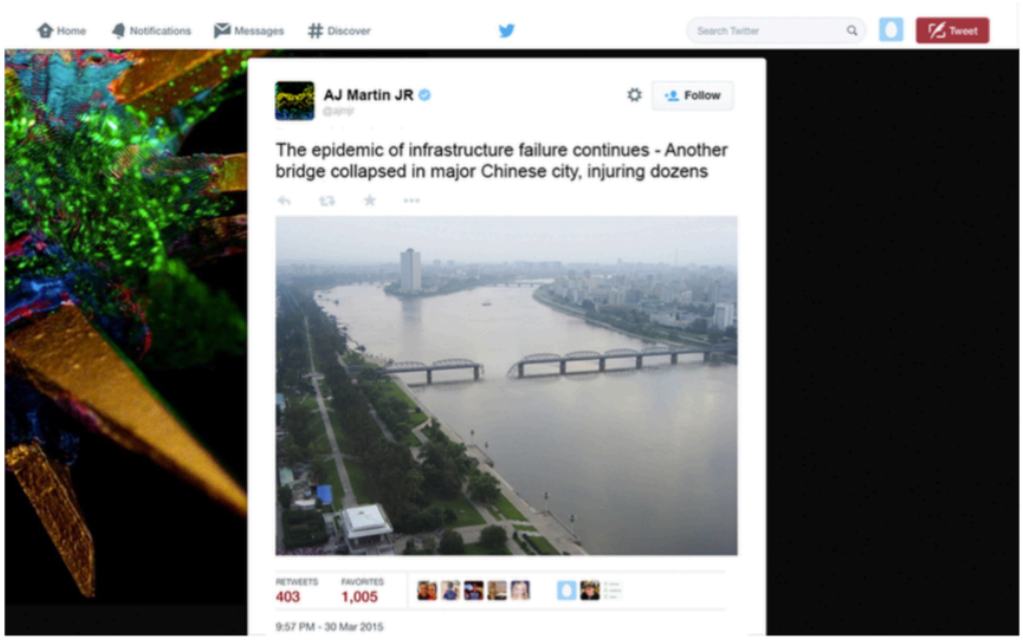
Preliminary analysis of our focus group results suggests that people generally perform poorly at making credibility assessments of online images. Further, non-image factors, such as the source of the image and its accompanying story, appear to play a much more significant role in participants' credibility judgment than image-specific factors such as inconsistencies in lighting and shadows. Contrary to our original expectations, when confronted with an implausible image, most subjects did not immediately suspect some form of image manipulation.

These findings have important implications. For image creators and publishers, the most productive ways to increase credibility ratings lie elsewhere, namely, in non-image features such as the source, story content, and online media interface. For example, images posted on Twitter can be perceived credible if they have a large number of retweets, favorites and followers, regardless of the actual image content.

For image consumers at large, it is advisable not to assume every image on the web is authentic.

This pilot study was limited by its small sample size and lack of demographic diversity. Moreover, the fake images analyzed in the study provided limited combinations of image content, source, and other contextual factors, making it difficult to isolate each factor's specific effect on image credibility perceptions. But these limitations will be addressed in the next stage of our work, a larger-scale online experiment on Amazon Mechanical Turk.

Sample Composition



Example 1: Mockup of atwitter post from a fake Twitter user, AJMartin JR, who is allegedly a verified member. The image shows a collapsed bridge which the post explains due to infrastructure failure in China. The source image shows a functioning bridge over the Taedong River in Pyongyang, North Korea. The manipulated image retouched and eliminated a section of the bridge.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1444840. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Goolsby, R. "On Cybersecurity, Crowdsourcing, and Social Cyber-Attack." Technical Report. 2013.

Ringel Morris, Meredith; Counts, Scott; Hoff, Aaron; Roseway, Asta; Schwarz, Julia. "Tweeting is Believing? Understanding Microblog Credibility Perceptions," Microsoft Research, Proceedings of CSCW 2012, February 2012.